Reproducibility and Data Integration in High-Dimensional Statistics

with Applications in Molecular Cancer Biology

Bilgrau, Anders Ellern

DOI (link to publication from Publisher):
10.5278/VBN.PHD.ENGSCI.00083

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

REPRODUCIBILITY AND DATA INTEGRATION IN HIGH-DIMENSIONAL STATISTICS

WITH APPLICATIONS IN MOLECULAR CANCER BIOLOGY

BY
ANDERS ELLERN BILGRAU

DISSERTATION SUBMITTED 2015

AALBORG UNIVERSITY
DENMARK

AALBORG UNIVERSITY
DENMARK

Reproducibility and Data Integration in High-Dimensional Statistics

with Applications in Molecular Cancer Biology

PhD dissertation

Anders Ellern Bilgrau

Dissertation submitted September, 2015

Thesis submitted: September 17th, 2015

PhD supervisors: Prof. Martin Bøgsted
Dept. of Clinical Medicine,
Aalborg University Hospital

Assoc. Prof. Poul Svante Eriksen
Dept. of Mathematical Sciences,
Aalborg University

PhD committee: Professor Jesper Møller (chairman)
Department of Mathematical Sciences
Aalborg University

Assistant Professor Kasper Daniel Hansen
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Professor Niels Richard Hansen
Department of Mathematical Sciences
University of Copenhagen

PhD Series: Faculty of Engineering and Science, Aalborg University

ISSN (online): 2246-1248
ISBN (online): 978-87-7112-366-1

Published by:
Aalborg University Press
Skjernvej 4A, 2nd floor
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2015

Preface

This dissertation conjoins and summarizes research carried out during my joint PhD fellowship at the Department of Mathematical Sciences, Aalborg University, and the Department of Haematology, Aalborg University Hospital. It concludes a multidisciplinary effort in biostatistics, bioinformatics, and computer science applied to molecular cancer biology and clinical research.

The thesis concerns statistical methods investigating high-dimensional genomic data and their implementation in statistical software. The methods were primarily developed with applications in genetics of haematologic cancers in mind but are suited for many other problems. As high-dimensional genomics entails a number of statistical and practical speed-bumps, the thesis has a strong focus on reproducibility and data integration of multiple experiments.

The thesis is primarily intended for (bio)statisticians. Some preliminary background in molecular and cancer biology is therefore provided for uninitiated readers. Along with this background material, this thesis consists of a collection of five scientific papers and three statistical software packages for the programming language R. All the papers are self-contained with separate bibliographies and numbering of sections, figures, tables, and equations. A short preface giving further context and extended discussion is given before each paper and to each software package.

There are numerous people that I need to acknowledge. First and foremost, I wish to express my sincere gratitude to my supervisors Martin Bøgsted and Poul Svante Eriksen for their pleasurable guidance and commitment to the project. You have been far to lenient with my unannounced office visits; I do not recall a single instance of you being too busy to answer my trivial questions. My office mate Steffen Falgreen deserves special thanks for helpful suggestions, comments, and friendship. As does Torben Tvedebrink, Mikkel Meyer Andersen, Julie Støve Bødker, Sara Correia Marques, and Lasse Hjort Jakobsen. Next, Hans Johnsen, Karen Dybkær, and Søren Højsgaard also deserve explicit recognition for giving me the opportunity. I thank also the rest of my colleagues at the Department of Mathematics and Department of Haematology for their help in various matters.

Thanks to Carel F. W. Peeters and Wessel N. van Wieringen for my two excellent and enjoyable stays in Amsterdam at the Department of Biostatistics & Epidemiology, VU University Medical Center, and the Department of Mathematics, VU University. The stays were gratefully supported by the Danish Cancer Society and the host institutions.

Anders Ellern Bilgrau
Aalborg, September 17th, 2015

To my friends and family.
And to The National, for *All The Wine*.

Thesis Details

Thesis Title:	Reproducibility and Data Integration in High-Dimensional Statistics—with Applications in Molecular Cancer Biology
PhD Student:	Anders Ellern Bilgrau
Supervisors:	Prof. Martin Bøgsted, Aalborg University Hospital Assoc. Prof. Poul Svante Eriksen, Aalborg University

The main body of this thesis consist of the following five papers:

[I] A.E. Bilgrau *et al.* (2015) *"Unsupervised Clustering and Meta-Analysis using Gaussian Mixture Copula Models"*, Accepted for Journal of Statistical Software

[II] S.M. Rasmussen*, A.E. Bilgrau*, *et al.* (2015) *"Stable Phenotype of B-Cell Subsets Following Cryopreservation and Thawing of Normal Human Lymphocytes Stored in a Tissue Biobank"*, Cytometry Part B: Clinical Cytometry, vol. 88, no. 1

[III] A.E. Bilgrau *et al.* (2015) *"Estimation of a Common Covariance Matrix for Multiple Classes With Applications In Meta- And Discriminant Analysis"*, Submitted to Annals of Applied Statistics

[IV] A.E. Bilgrau*, C.F.W. Peeters*, *et al.* (2015) *"Targeted Fused Ridge Estimation of Multiple Inverse Class Covariance Matrices from High-Dimensional Data"*, Submitted to Journal of Machine Learning Research

[V] A.E. Bilgrau *et al.* (2015) *"Unaccounted Uncertainty from qPCR Efficiency Estimates Imply Uncontrolled False Positive Rates"*, Submitted to Nucleic Acids Research

Along with the five papers, the thesis contains the following three software packages developed for the statistical programming language R:

[I] A.E. Bilgrau, *et al.*, **GMCM**: Fast Estimation of Gaussian Mixture Copula Models. R package v1.2.1, `https://cran.r-project.org/package=GMCM`

[II] A.E. Bilgrau, *et al.*, **DLBCLdata**: Diffuse Large B-Cell Lymphoma data. R package v0.9, `https://github.com/AEBilgrau/DLBCLdata`

[III] C.F.W. Peeters, A.E. Bilgrau, and W.N. van Wieringen **rags2ridges**: Ridge Estimation of Precision Matrices from High-Dimensional Data. R package v2.0, `https://cran.r-project.org/package=rags2ridges`

In addition to the main papers and software, the following manuscripts were also published, although not considered part of this thesis:

[1] K. Dybkær*, M. Bøgsted*, S. Falgreen, J.S. Bødker, M.K. Kjeldsen, A. Schmitz, A.E. Bilgrau, *et al.* (2015) *"A Diffuse Large B-Cell Lymphoma Classification System That Associates Normal B-Cell Subset Phenotypes with Prognosis"*, Journal of Clinical Oncology, vol. 33, no. 12: pp. 1379–1388

[2] T.C. El-Galaly, A.E. Bilgrau, *et al.* (2015) *"A Population-Based Study of Prognosis in Advanced Stage Follicular Lymphoma Managed by Watch and Wait"*, British Journal of Haematology, vol. 69, no. 3: pp. 435–444

[3] M. Bøgsted, A.E. Bilgrau, *et al.* (2013) *"Proof of the Concept to Use a Malignant B Cell Line Drug Screen Strategy for Identification and Weight of Melphalan Resistance Genes in Multiple Myeloma"*, PLoS ONE, vol. 8, no. 12: e83252

[4] K.S. Bergkvist, M. Nyegaard, M. Bøgsted, A. Schmitz, J.S. Bødker, S.M. Rasmussen, M. Perez-Andres, S. Falgreen, A.E. Bilgrau, *et al.* (2014) *"Validation and Implementation of a Method for Microarray Gene Expression Profiling of Minor B-cell Subpopulations in Man"*, BMC Immunology, vol. 15, no. 3

[5] V. Kurande, A.E. Bilgrau, *et al.* (2013) *"Interrater Reliability of Diagnostic Methods in Traditional Indian Ayurvedic Medicine"*, Evidence-Based Complementary and Alternative Medicine, vol. 2013, no. 658275

[6] M. Grønkjær, C.F. Hasselgren, A.S.L. Østergaard, P. Johansen, J. Korup, M. Bøgsted, A.E. Bilgrau, P. Jensen (2015) *"Bone Marrow Aspiration: A Randomized Controlled Trial Assessing the Quality of the Bone Marrow Specimen using Slow and Rapid Aspiration Techniques and Evaluation of the Pain Intensity"*, Accepted for Acta Haematologica

[7] T.C. El-Galaly*, A.E. Bilgrau*, *et al.* (2015) *"Circulating tumor necrosis factor-α and YKL-40 level is associated with remission status following salvage therapy in relapsed non-Hodgkin lymphoma"*, Leukemia & Lymphoma, Early Online

A star (*) denotes shared first authorship. This thesis has been submitted for assessment in partial fulfillment of the PhD degree. The thesis is based on the submitted or published scientific papers which are listed above. Parts of the papers are used directly or indirectly in the extended summary of the thesis. As part of the assessment, co-author statements have been made available to the assessment committee and are also available at the Faculty. The thesis is not in its present form acceptable for open publication but only in limited and closed circulation as copyright may not be ensured.

Abstract

High-dimensional data and statistics permeate modern science and technology. The dimension of data can be loosely defined as *the number of measurements within each observation* and is considered 'high' when it exceeds the number of observations. Today's biotechnology can measure millions of genetic markers and provides an unprecedented detailed view of human diseases. In clinical research, such high-dimensional data has brought a yet unfulfilled promise of personalized medicine, tailor-made drugs, and clinical decisions based on our genetic makeup and history. The analysis of these data-intensive experiments presents some of the biggest challenges in the field of statistics.

From a statistical perspective, high-dimensionality is accompanied by a host of theoretical, technological, and practical problems and pitfalls, which may lead to spurious findings, irreproducible results, and invalid conclusions for careless researchers. Indeed, concerns of irreproducible findings have been raised in high-dimensional genetics and the overlap in findings between studies have been described as *disappointingly small*. This thesis explores various aspects of (bio)statistics related to the reproducibility of genetic findings.

A straightforward way of increasing reproducibility is integrating data from multiple experiments. This is desirable as genetic studies are costly and often have small sample sizes. Following this, Paper I discusses a previously proposed model for integrating data from multiple studies and quantifying the degree of reproducibility. The paper improves upon a more general model and provides a substantially faster software implementation. Paper II assesses the impact on genetic profiling when cryopreserving cells. Cryopreservation is an important and convenient tool for storing cells which assumes that findings will reproduce if fresh cells are used instead.

One proposed explanation for the poor reproducibility are the common gene-by-gene screenings that, too simplistically, consider genes independent. Instead, genes are highly *dependent*, operating in tightly regulated networks. However, statistical network analysis is known to be highly unstable and thus often irreproducible. To this end, Papers III and IV propose methods for identifying gene networks by aggregating information across multiple datasets. This is feasible as genetic data often is made publicly available in large online repositories. Combining data from multiple datasets may increase the sensitivity and stability of the estimated networks yielding more reproducible results. Again, fast and easy-to-use software implementations are freely available.

The poor reproducibility has partly been acknowledged as genetic screenings are often called *hypothesis generating*, implying a need for validating findings in independent experiments. However, as discussed in Paper V, the analysis of the gold-standard validation experiments can be overly optimistic; a highly undesirable property for *validation*.

Dansk Resumé

Høj-dimensionel data og statistik gennemsyrer moderne videnskab og teknologi. Dimensionen af data kan løst defineres som *antallet af målinger i hver observation* og den anses for at være 'høj', når den overstiger antallet af observationer. Moderne bioteknologi kan måle millioner af genetiske markører, der giver et hidtil uset detaljeret billede af sygdomme. I klinisk forskning har høj-dimensionelle data bragt et endnu uopfyldt løfte om personlig medicin, skræddersyede lægemidler og kliniske beslutninger, baseret på vores genetiske sammensætning og historie. Analysen af sådanne data-intensive eksperimenter hører til nogle af de største udfordringer inden for statistik.

Fra et statistisk synspunkt er høj-dimensionel data ledsaget af et væld af teoretiske, teknologiske og praktiske problemer og faldgruber, som kan føre til falske fund, ikke-reproducerbare resultater, og ugyldige konklusioner for uforsigtige forskere. Faktisk er bekymringer om ikke-reproducerbare fund blevet rejst i høj-dimensionel genetik, og de overlappende resultater mellem undersøgelser er blevet beskrevet som *skuffende få*. Denne afhandling omhandler forskellige aspekter af (bio)statistik og reproducerbarheden af genetiske studier.

En simpel måde at øge reproducerbarheden på er at integrere data fra flere eksperimenter. Dette er hensigtsmæssigt, da genetiske undersøgelser er dyre og ofte har små stikprøvestørrelser. Med hensyn til dette diskuterer Artikel I en model, der kan integrere data fra flere studier samt kvantificere graden af reproducerbarhed. Artiklen præsenterer en mere generel model og en væsentligt hurtigere softwareimplementering. Artikel II vurderer indvirkningen på de genetiske profiler, når celler kryopræserveres. Kryopræservering er et vigtigt og praktisk værktøj til opbevaring af celler, under forudsætningen at resultaterne vil kunne reproduceres, hvis der anvendes friske celler i stedet.

Én foreslået forklaring på den dårlige reproducerbarhed er de almindelige gen-for-gen screeninger, som alt for forsimplet antager at gener er uafhængige. Gener er i stedet yderst *afhængige* og opererer i tæt regulerede netværk. Statistisk netværksanalyser er imidlertid meget ustabile og er derfor ofte ikke reproducerbare. Til dette formål foreslår Artikel III og IV metoder til at identificere gen-netværk ved, at samle oplysninger på tværs af flere datasæt. Dette er muligt, da genetiske data ofte gøres offentligt tilgængelige i store online databaser. Kombinationen af flere datasæt kan øge følsomheden og stabiliteten af de estimerede netværk, og dermed også reproducerbareheden af resultaterne. Også her er hurtigt og nemt software frit tilgængeligt.

Den dårlige reproducerbarhed er delvist blevet anerkendt, da de genetiske screeninger ofte kaldes *hypotesegenererende*. Dette indebærer et behov for validering af resultaterne i uafhængige forsøg, men som omtalt i Artikel V, kan analysen af 'guldstandard' valideringseksperimenter være alt for optimistiske; en særdeles uønsket egenskab i *validering*.

Contents

Part A

Background

Introduction

The completion of the Human Genome Project in 2003 highlights a new era in biology and clinical medicine with vast amounts of data [31, 64]. More than a decade earlier, the project set out to identify the sequence of our roughly 3.3 billion base pairs of DNA and 20 thousand genes; the human genome. The decade long $3 billion effort, generated some 500 gigabytes of biological data. Today, the same feat can be accomplished in a matter of days for less than $10,000 in the race towards the nicknamed *$100-genome*. These biotechnological advances of rapid data collection, dubbed *high-throughput*, has evolved from genomics into a wide range of *omics*-fields. Genomics, transcriptomics, proteomics, metabolomics, lipidomics, all provide unprecedented detail of the remarkable molecular machinery of life and invaluable information about human evolution, development, physiology, and effective medicine.

High-throughput biotechnology allow researchers to perform genome-wide searches and speed up the traditional slow research process. A tremendous focus has been put on searching for genetic markers amongst the large number of candidates to classify diseases, understand the pathogenesis, and provide a prognosis or even a prediction of the outcome of the disease. This has brought a still unfulfilled promise of personalized medicine that can offer tailor-made clinical decisions and drugs based on the genetic makeup of the patient [4].

However, the scientific principle of reproducible research has been somewhat neglected and concerns have been raised of non-reproducible genetic findings and studies [4, 18, 29, 30]. Several possible explanations exist for the poor reproducibility. High-throughput data is accompanied by a host of technological, experimental, and theoretical problems as well as the practical ones, all of which will substantially deteriorate or destroy the reproducibility when aggregated.

Technological and experimental issues include a tremendous number of laboratory variables, disparate technological platforms, and artifacts hereof. From the statistical viewpoint, high-throughput data has many theoretical properties working against reproducibility and researchers. Added to this mix comes cumbersome handling of the experimental data with plentiful choices in preprocessing and analyzing the data. All these lessen the expected reproducibility, as the heterogeneity of the experiments increases. Moreover is the inherently difficult and complex biological systems studied which are hard to control as indicated in the following section. For careless researchers these issues—though separately minor—often imply spurious findings, ultimately irreproducible results, and possibly invalid conclusions.

It is worth noting that high-throughput technologies are not unique to biology. Nearly all fields, including chemistry, economics, astronomy, physics, forensics, and computer science, have increasingly made use of the fashionable

'big data' and its data-intensive applications. High-dimensional data and its problems are ubiquitous in all of modern science, technology, and statistics.

1 Molecular cancer biology and biotechnology

This section establishes basic preliminary knowledge and terminology of molecular biology necessary for understanding the DNA microarray technology used throughout the thesis. The section also provides some additional subject-matter context to the statistical methods considered in the papers.

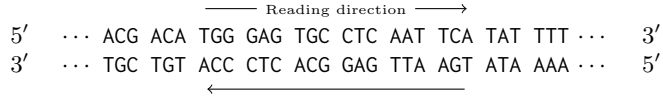
1.1 Nucleic acids, proteins, and genes

Within the nucleus of each cell resides a full copy of the schematics and complete set of instructions for its development and function. This information is contained within our DNA (deoxyribonucleic acid) which are very large polymers consisting of contiguous *nucleotides* as illustrated in Figure 1. The instructions are thus encoded linearly in millions of DNA nucleotides, wound into the double helix, and tightly organized in large structures called chromosomes. Each human somatic cell contains 23 pairs of such chromosomes amounting to 46 molecules of DNA.

The DNA backbone is two chains of alternating phosphate groups and *deoxyribose* sugars. Attached to this backbone are so-called *bases*, which comes in the four chemical flavors of guanine (G), adenine (A), cytosine (C), and thymine (T). The sequence of the bases carries our genetic information as the backbone of the DNA is the same everywhere. DNA is double-stranded where a given base of one strand is always paired with its complementary base of the other; A is complementary to T while C is complementary to G [42]. As the information-storing molecule, DNA is stable and long-lived.

RNA (ribonucleic acid), on the other hand, is much like a single stranded version of DNA with a markedly shorter life. RNA is distinguished from DNA by using ribose instead of deoxyribose as the sugar in its backbone and the base uracil (U) replaces T [42]. As will be apparent, the RNA is the executive molecule of the body.

For reasons not explained, DNA and RNA are directional with a so-called 5′-end (read 5-prime-end) and 3′-end. This directionality is important as the enzymes that read, transcribe, and translate the genetic information do so in the $5' \to 3'$ direction. The two complementary strings of DNA run in reversed directions, and the terms *up-stream* and *down-stream* refer to bases toward the 3′- and 5′-ends, respectively. In essence, DNA can be illustrated by two strings of the letters representing the *base pairs*;

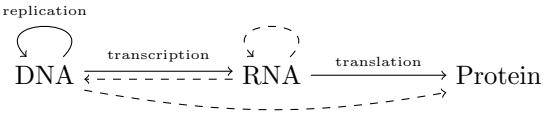


Note, that only one strand is needed for all the genetic information.

For particular stretches of the DNA, each three-base sequence serve as codes, called codons, for particular amino acids. Thus, DNA directly encodes proteins, which are polymer chains of amino acids. A gene can be defined as region of DNA from which functional RNA or protein can be created. As such, it is the basic unit of heredity. The length of a gene is typically in the range of 1,500–100,000 base pairs, although, genes that are two million base pairs long are known to exist. The upper half of Figure 1 illustrates these relationships between cells, chromosomes, DNA, and genes.

1.2 Central dogma of molecular biology

The central dogma of molecular biology describes the principal flow of the genetic information encoded in our DNA into working proteins [10, 11]. In its simplest and most widely known form by Watson et al. [66], DNA can be replicated or it can be transcribed into RNA and RNA can be translated into protein as depicted by the solid arrows below:



DNA replication is essential for cell division to provide cells with identical copies of all chromosomes. Transcription and translation are the basic processes converting the genetic information in the DNA into functional RNA molecules or proteins. More precisely, transcription is the process of transcribing DNA into precursor messenger RNA (pre-mRNA) molecules called *transcripts*. Besides proteins, genes can also code for functional RNA molecules such as transfer RNA (tRNA) and ribosomal RNA (rRNA), both used in the translation process. The three enzymes RNA-polymerase I, II, and III transcribe the rRNA, pre-mRNA, and tRNA, respectively. A *splicing* process then converts pre-mRNA into mRNA, which can leave the nucleus and wander to the ribosomes. In the ribosomes, the codons of the mRNA is translated to the amino acid chain that makes up the protein using tRNA and rRNA [42]. At last, the amino acid chain folds into the active functional protein. The molecules and steps from DNA to RNA to protein are seen illustrated in Figure 1.

However, as hinted by the dashed arrows above, many other processes exist which can alter how the standard transcription and translation takes place. Since Watson et al. [66] and Crick et al. [10] the central dogma has been extended and often contradicted. Today, many known processes such as reverse transcription, (alternative) splicing, the existence of alternative transcription start-sites, introns, non-coding RNAs, and epigenetic mechanisms such as DNA methylation and histone modification have been realized and the view above is now regarded as an over-simplified yet useful model [22, 25, 39]. Long non-coding RNAs and microRNAs (miRNA) are currently highly researched examples of transcripts of DNA that does not encode protein but does possess regulatory functions [39]. Alternative splicing enables a single gene to produce

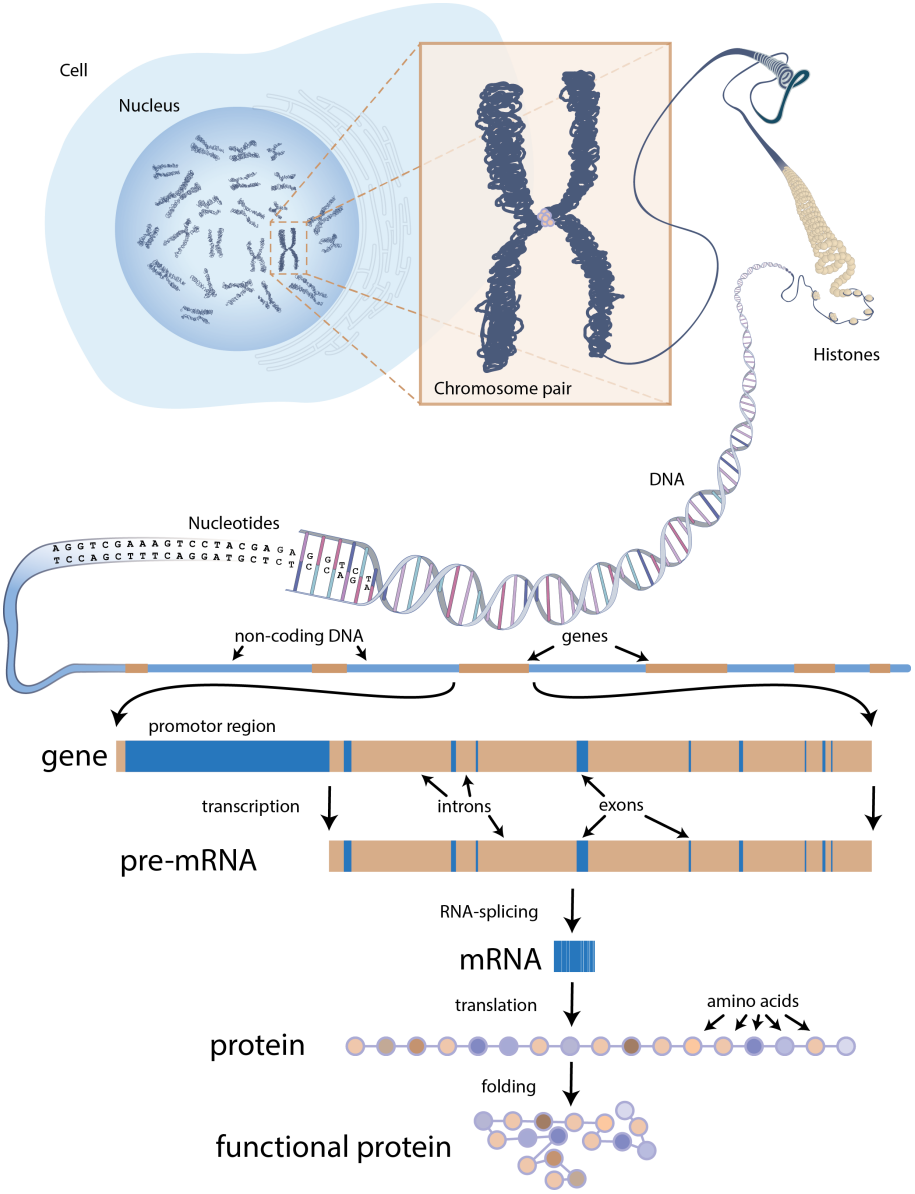


Fig. 1: Illustration of the structure of the chromosomes, DNA, RNA, and the central dogma of molecular biology. Note, that the chromosomal structures depicted are only present near cell division in the so-called meta-phase.

many different proteins which immensely expands the variety of manufactured proteins to the order of two million from our approximately 23 thousand genes [31, 64].

1.3 Gene activity and expression

The expression of a gene can be loosely defined as the degree to which it manifests itself. In this sense, the expression of a gene is almost synonymous to its activity. Genes that are highly active produce many mRNA transcripts that yield more translated protein, and causes the gene to be highly expressed. Conversely, genes which are (nearly) inactive produce very few or no transcripts and thus have a low expression. Relative to healthy tissue, the expression of a gene can therefore be high or low in cancer tissue, for example. The identification of such up- or down-regulation of gene expressions is an archetypal exercise in molecular cancer biology and of central interest as they hint at the pathogenesis. Section 2.1 provides a brief review of this task.

However, the expression of a gene depends on an enormous set of factors as alluded above. At each stage from DNA to the gene product, there exists a multitude of other gene products that, directly or indirectly, control and regulate the processes [50]. Such regulation might be activation or inhibition of genes. Micro-RNAs are one example of functional non-coding RNA which can inhibit the translation process and thereby turning off the gene expression. Other examples of regulation are alteration of the mRNA degradation speed and modification of the affinity for specific alternative pre-mRNA splicing variants. Since such regulation mechanisms are carried out by molecules from other genes and *their* expression, all gene expressions are conducted through a tightly regulated concert to carry out their functions. A manifestation of this complex network of gene interactions has perhaps been seen by poor concordance of mRNA and protein levels [14, 37]. Schwanhäusser et al. [50], however, ascribed the poor correlation to technical difficulties and concluded that protein abundance is predominantly controlled at the translation level. In any case, it is unsurprising that gene expressions exhibit a complex network of dependencies. Papers III and IV of this thesis presents statistical models attempting to elucidate these gene networks. Section 1.5 explains how the gene activity and expression are routinely measured by the *DNA microarray* technology.

1.4 Diffuse large B-cell lymphoma

Cancer is characterized by abnormal cell growth that aggressively spreads and invades other parts of the body. Unlike benign tumors, the intrusive cancerous cells often spread through the blood and lymphoid systems and metastasize where they again proliferate. The uncontrolled growth destroys and kills neighboring tissue, organs, and ultimately the patient if left unchecked.

It is believed that cancer arises from a single cell with one or more genetic changes that cause uncontrolled cell division and proliferation. Such a progres-

sively damaged cell may acquire sufficient genetic changes and dysregulation as to, e.g., disable or circumvent the pathway of apoptosis; the (healthy) process of programmed cell death.

Approximately 90% of lymphomas are non-Hodgkin's lymphomas of which the largest subgroup is diffuse large B-cell lymphomas (DLBCL) [28, 58]. Two major molecular subgroups of DLBCL has been identified by Alizadeh et al. [1] named activated B-cell-like (ABC) and germinal center B-cell-like (GCB) after their resemblance in gene expression profiles with these cells. Patients of the GCB type show a clinically significant favorable overall survival compared those of the ABC type [1, 48]. Although these DLBCL subtypes were identified more than a decade ago, they are still treated as a singular disease in clinical practice. Differentiated treatments have only recently started to appear in clinical trials [41, 49].

Besides many associative differences, relatively few biological and functional differences between the subtypes are known. Some of the known oncogenic mechanisms known to distinguish the subtypes include recurrent t(14:18) translocation in GCBs, trisomy 3 in ABCs, deletion of the inhibitor of kinase 4A-alternative reading frame (INK4A/ARF) locus, and activation of the anti-apoptotic NF-κB signaling pathway [34, 58].

I remark here, that patients who are classified as neither ABC nor GCB have inconsistently been deemed unclassified (UC) [e.g. 65] or type III [e.g. 48] in the literature. Discussions as to whether a third type really exists or not are ongoing. However, in the statistical classification system of the subtypes by e.g. Wright et al. [67], using a naïve Bayes classifier, a distinct third subtype is meaningless by construction; the only sensible name here is UC. In this setup, samples are assumed either ABC or GCB, and failing to be classified as either does not necessarily imply a third type. The classification scheme by [8] used in Paper IV has three subtypes by construction. Each sample is given a probability of belonging to ABC, GCB, and Type III and the class corresponding to the largest probability is selected. This raises a slippery slope of questions of a fourth subtype if neither probability is much larger than 1/3.

1.5 Gene expression profiling by DNA microarray

Today's mass-market *DNA microarrays* simultaneously measures the gene expressions of thousands to millions of gene sub-units. They are simple, powerful, and comprehensive tools for systematically exploring the genome [7]. Manufacturers claim the microarrays to be whole-genome as they investigate nearly all human genes.

As illustrated by Figure 2, a microarray is a silicon chip on which *probes* are placed in a rectangular grid of *spots*. Each spot is a collection of custom designed *probes*—so-called *oligonucleotides* of single stranded complementary DNA—that investigate specific DNA fragments of interest.

To quantify the gene expressions using microarrays, the sample of a cancerous tumor, say, is first prepared by purifying the mRNA. The prepared sample

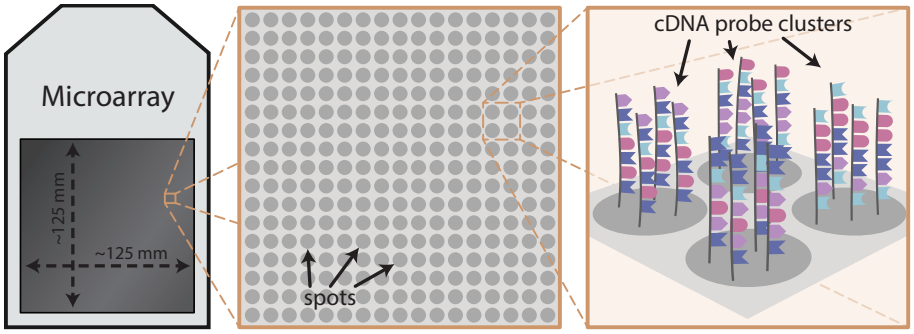


Fig. 2: Simplified illustration of a microarray. From left to right: A microarray and approximate dimensions of the chip. Magnification of the array, showing a 17×17 area of spots of clusters of replicated single stranded complementary DNA. Magnification of four spots and showing the nucleotides of each single stranded complementary DNA in each cluster.

is then converted into *tags* by first fragmenting the mRNA, breaking it into smaller pieces, subsequently converting the mRNA fragments into DNA by reverse transcriptase, and lastly 'labeling' the DNA by appending fluorescent molecules. The labeled tags are then applied to the microarray where the tags complementary to probes are allowed to hybridize. After this hybridization the excess remaining DNA material is rinsed off and the gene expressions are ready to be measured.

The central premise of microarrays is that the amount of material hybridized to the probes is proportional to the number of transcripts in the original samples; a very active gene is expected to produce a high number of mRNA transcripts, and thus, many tags are expected in the prepared sample. The number of tags attached to each probe then provides a surrogate measure of the relative gene expressions. The actual measurements are made by capturing the fluorescence intensity of each probe with an image while illuminating the microarray using laser light. The expression level is then be obtainable as the reemission intensity reflects the abundance of hybridized tags at each spot. Figure 3 shows an such an image and Figure 4 illustrates the conceptual molecular workings explained above. Using image processing, the expression of a particular gene is derived from the pixel intensities, e.g. of Figure 3, corresponding to the known locations of the probes. As easily imagined, the procedure and data quality depends on a huge set of factors including sample purity, efficiency of labeling and RNA purification, non-uniform images, etc., all of which are unwanted sources of noise and variation. The set of obtained expression estimates for a given sample is called a *gene expression profile* (GEP).

The microarrays used in this thesis are manufactured by Affymetrix by lithographically directly synthesizing the probes onto the chip. They are manufactured by sequentially assembling the probes using masks in an bottom-up fashion. The microarray is first masked and the partly exposed chip is then covered in a solution of, for instance, G nucleotides. The G nucleotides are then

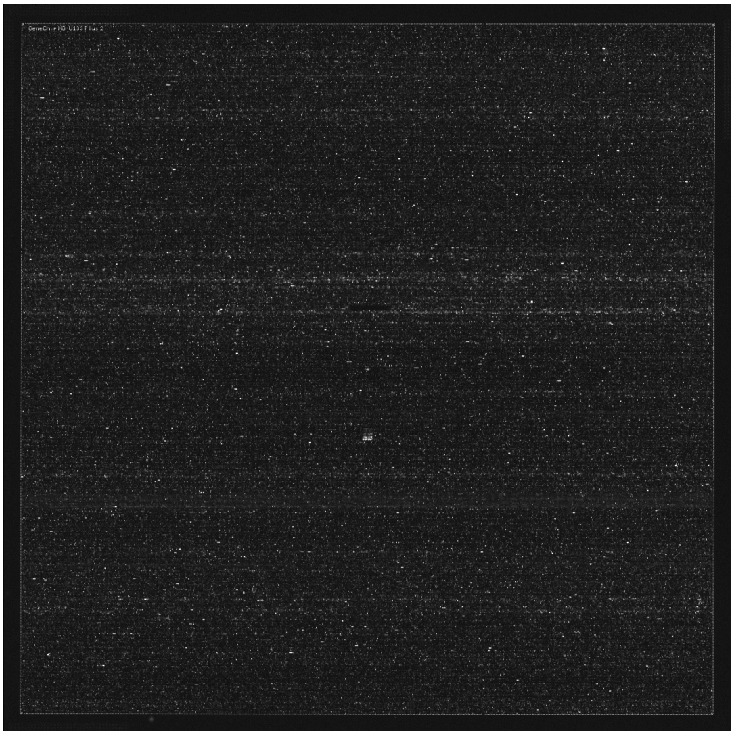


Fig. 3: An image produced by the Affymetrix GeneChip HG-U133 plus 2.0. Each dot corresponds to a single spot of probes. There are approximately 1,355,000 spots on the microarray.

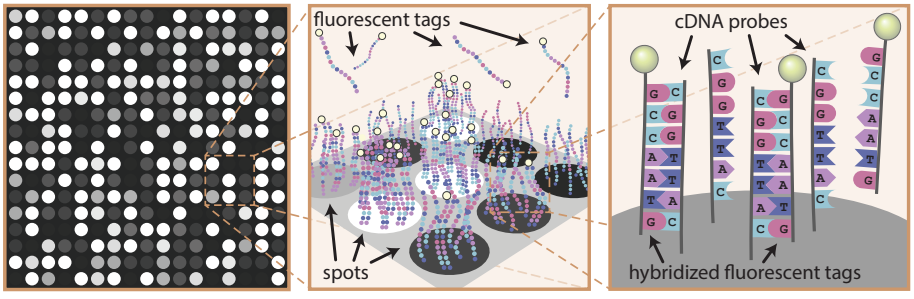


Fig. 4: Illustration of an illuminated microarray. From left to right. Part of the illuminated microarray. Magnification of nine spots showing the clusters of complementary DNA that make them up, some hybridized with the florescent tags from the sample. Magnification of a single simplified spot and the nucleotides of each single stranded cDNA. The fluorescent tags from the samples hybridize to the probes of the spot if they are complementary.

synthesized using light onto the starting anchor or previous partial DNA chains in the locations specified by the mask. Next, the microarray is rinsed to remove excess `G` nucleotides and the whole process is repeated, building the probes in parallel until the desired sequences are constructed.

In compromise between production costs, probe uniqueness, and binding strength, the probe lengths on both the *GeneChip Human Exon* 1.0 *ST* (exon) and *GeneChip Human Genome U133 plus* 2.0 (U133) arrays are 25 nucleotides. Short or less unique probes and low binding strength allows tags to hybridize to probes that are partially matching. Probes mismatching at the middle nucleotide have been introduced on the U133 chip to gauge this so-called *cross-hybridization*. From the probe-pair of the perfectly matching probe (PM) and the corresponding mismatching probe (MM), the degree of non-specific binding can be estimated. The use of the MM probes, however, has stopped, as they seem to introduce more noise than they remove [68]. For example, MM probes often perplexingly shows a higher expression than their PM counterpart.

The probe pairs are organized into probe sets, each of which investigate particular a gene or gene sub-unit. There are approximately 1.3 million probes on the U133 array, which are arranged into *probe sets* of 11 probe pairs. On average, the U133 array has eleven probes for each probe set which amounts to some 55 thousand probe sets. In contrast, the exon array consists of 6.5 million probes with four probes per probe set on average [52, 55].

Preprocessing is the extraction of the raw image data and conversion to biologically meaningful data. It consists of image processing and a statistical model summarizing the probe sets into gene expression levels. Simultaneously, preprocessing aims to remove much of the introduced noise [68].

As introductory mentioned, DNA sequencing platforms are increasingly popular. Such platforms works by determining the sequence of hundreds of millions of fragments of DNA in parallel. Here, the estimated expression level is the number of times a given sequence is observed. In addition and contrary to microarrays, no prior knowledge about the specific sequences to investigate is required for these technologies. Although sequencing certainly looks to be the future, the usage of microarrays is still well established and widespread following Figure 5.

2 Statistical analysis of microarray data

This section covers two broad approaches to the analysis of microarray data that are considered in this thesis.

2.1 Differential gene expression

The perhaps most archetypal analysis of gene expression profiles is the so-called analysis of differential gene expression [17]. Searching PubMed.gov for '`differential expression analysis`' yields some 43 thousand articles as of

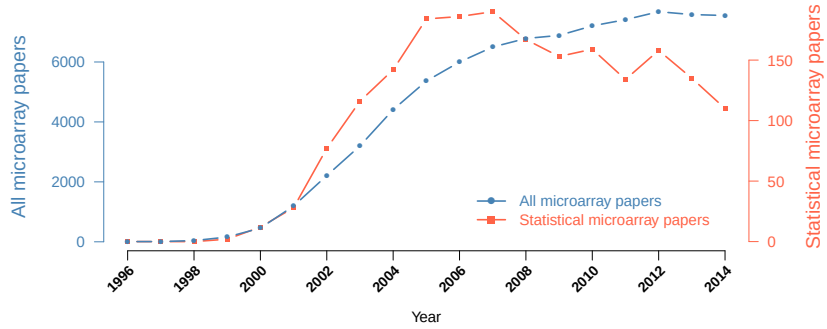


Fig. 5: The number of publications each year using microarray and the statistical analysis hereof as found by PubMed.gov. The searches were carried out using the queries '`(microarray* OR gene-chip* OR genechip*)`' and '`(microarray* OR gene-chip* OR genechip*) AND (statistical method* OR statistical techniq* OR statistical approach* OR statistical analy* or statistical pro*)`' for the categories *all microarray papers* and *statistical microarray papers*, respectively. This is, essentially, an updated plot of Mehta et al. [40] attempting to not count sequencing papers.

August, 2015. In the canonical example, gene expression profiles are collected for a case (e.g. treatment) and a control (non-treated) group. The genes of research interest then answers the question: *'Which genes are changed due to the treatment?'* Usually, each gene is investigated individually and independently by testing the statistical hypothesis of no mean difference in gene expression. This allows researchers using microarrays to quickly screen all human genes for differential expression and provides a list of genes ordered by relevance.

The naïve approach to this problem employs gene-by-gene t-tests therefore computing t-statistics

$$t_g = \frac{\hat{\mu}_g^{\text{case}} - \hat{\mu}_g^{\text{ctrl}}}{\text{se}_g}, \tag{1}$$

for each gene g where $\hat{\mu}_g^j$ is the estimated expression level of gene g in group j and se_g is the standard error of the numerator.

By thresholding the corresponding p-values, the most significant genes can be considered. The first problem of this methodology of testing thousands to millions of statistical hypotheses is that it considerably increases the probability of sporadic false findings due to the so-called multiple hypothesis testing problem. This problem often riddles the detected biomarkers with false positives. To combat this, p-value corrections are employed in an attempt to control the false positive rate or false discovery rate. The two most popular corrections for adjusting the p-values are the Bonferroni or Benjamini-Hochberg corrections [5] though many exists. Multiple hypotheses testing is an inevitable part of analyzing high dimensional data.

Other properties in addition to the multiple testing problem make the analysis of microarray data less trivial. For instance, the signal-to-noise ratio increases with increasing expression level and the expression variance are often

gene-specific. Therefore, a less naïve approach is the *moderated* t-tests employed by using the statistic given by

$$t_g = \frac{\hat{\mu}_g^{\text{case}} - \hat{\mu}_g^{\text{ctrl}}}{\text{se}_g + s_0} \tag{2}$$

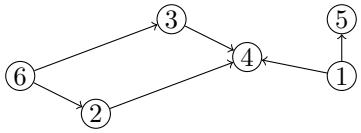
where s_0 is some positive moderating constant. Various choices of s_0 have been suggested [47, 53, 61]. Tusher et al. [61] uses the median of the distribution of standard deviations for all genes as moderation and derives p-values by permutation. Smyth [53] estimates s_0 in context of an empirical Bayesian model.

Differences in experimental settings and technology among different experiments also makes analysis, especially meta-analysis and data integration, of cancer microarray data challenging.

2.2 Network analysis of gene expression

The standard methodology of differential expression ignores the fundamental network structure of genes. Higher accuracy and precision in the findings can therefore potentially be achieved if genes are modeled as a network. In addition, such network modeling might yield much more specific evidence for causal regulatory pathways rather than the more associative results of differential expression analysis.

There is a large number of different approaches to the problem of network estimation. See e.g. Markowetz and Spang [38] for a review or the introductions of papers III and IV. Many of these fall under the category of so-called graphical models. Graphical models utilize *graphs* to specify the (in)dependence structure of the genes. Informally, a graph is a collection of *vertices* and *edges* as depicted below:



Formally, a graph $\mathcal{G}$ is an ordered pair $\mathcal{G} = (V, E)$ of sets where $V = \{1, \ldots, p\}$ is a set of p vertices and $E \subseteq \{(u, v) | u, v \in V\}$ is a set of edges which are pairs of elements of V. If the edges of E are all ordered, it is a *directed graph*. If the edges are unordered, then the graph is *undirected*. The graph above is directed and ignoring or omitting the arrows results in an undirected graph. In graphical models, a variable corresponds to a particular vertex and a dependency to an edge. Independencies then correspond to the absence of edges. However, the exact meaning of the edges vary widely and corresponds to different classes of models.

A graphical models is defined by the conditional (in)dependencies of the genes where the chosen scheme of conditioning give rise to different *classes*

of models and meaning of edges. More precisely, suppose the random vector $\mathbf{X} = (X_1, \ldots, X_p)$ corresponds to the vertices of $\mathcal{G}$. An edge is then defined to be absent between vertex i and j if and only if X_i is conditionally independent of X_j. Formally, the edge (i, j) is present if and only if

$$X_i \not\perp\!\!\!\perp X_j \mid \mathbf{X}_S$$

where S is some subset $S \subseteq V$. The defining property of the class of models considered is the choice of vertices S one conditions on.

If we condition on nothing, i.e. $S = \emptyset$, then marginal independence is considered. This leads to so-called correlation networks, relevance networks, or co-expression networks, where each edge correspond to a dependent gene-pair or, equivalently, a non-zero (*marginal*) correlation [26].

If $S = V \setminus \{i, j\}$ then so-called Markov random fields or Markov networks are considered. An edge in this case corresponds to a pair of genes which are dependent conditional on all remaining genes, or, equivalently have a non-zero partial correlation. Hence, edges are present whenever the question *'can the correlation between X_i and X_j be explained by* all *other genes?'* can be answered in the negative. If $\mathbf{X}$ is assumed to be multivariate Gaussian then the class of Gaussian graphical models (GGM) is considered [15, 32]. This is the class of models considered in papers III and IV.

Likewise, if $S = V \setminus \{k\}$ is used for all $k \neq i, j$, then edges are defined as the negative answers to *'can the correlation be explained by* any *other gene?'* Such a network is called a first order conditional network which generalizes easily to higher orders.

Lastly, if no restriction is put on S we arrive at Bayesian networks which asks *'can the correlation be explained by* any set *of other genes?'* Unlike the previous models, this has the added property that a direction often can be inferred but it requires the true underlying network to be acyclic.

Network analysis is however much more challenging than differential expression analysis. Results are harder to visualize, present, and interpret. In general, estimation of networks is much more variable and unstable than gene-by-gene analyses. One obvious way to obtain more stable results is to increase the sample size. However, large gene expression are costly with small samples sizes compared to the number of genes. We have therefore focused on gathering and integrating multiple datasets to obtain more stable network estimates. This strategy is possible as many gene expression datasets now are publicly available in large online data repositories. To this end, Papers III and IV describe models for estimating precision matrices, and thus GGMs, in multiple datasets or classes of data. Another interesting alternative method to this integration problem was proposed by Li et al. [36]. Their method searches for so-called *recurrent heavy subgraphs* among many co-expression networks.

Additional stability can potentially be achieved by using prior knowledge of gene networks also available in online databases. The method of Paper IV also allows for integrating such prior knowledge.

3 Reproducibility of genetic experiments

The reproducibility of experimental findings in high-throughput molecular biology have been disputed and challenged [18, 24, 29, 30]. Ein-Dor et al. [18] claimed that thousands of samples are needed to achieve the desired robustness and reliability of the gene list, in context of predicting clinical outcomes of breast cancers. They noted that while different research groups identify sets of genes with good prognostic performance, the number of common genes in these lists is 'disappointingly small'. The same seems true for even replicated experiments [59].

In contrast, Zhang et al. [70] argue that the lack of reproducibility is an apparent one. In high-dimensional setups, multiple hypothesis testing is usually employed to select candidates for further research. As such, the multiple testing problem superficially seems an important culprit that riddles the gene lists with false discoveries. However, Zhang et al. [70] showed that the false discovery rate (FDR) in separately determined lists can be very low. This suggests that various degrees of co-expression, induced or biological, between genes may be a principal cause for the poor number of overlapping genes. Co-expression introduces dependency between tests which make attempts to control the type I error rate very challenging in practice. The simple Bonferroni correction is nearly always too conservative. Frequently the same holds for the Benjamini-Hochberg corrections. This again motivates the more realistic modelling of genetic data through network analysis. Confer Zhang et al. [70] for a more in-depth discussion of the apparent lack of reproducibility.

3.1 Reproducibility and its flavors

The principle of reproducibility is considered a hallmark of science. Its importance was emphasized in the 17th century by Robert Boyle, but only later was it firmly established as part of the scientific method by Karl Popper and Ronald Fisher [21, 51]. Popper [43] wrote *'only when certain events recur in accordance with rules or regularities, as is the case with repeatable experiments, can our observations be tested—in principle—by anyone.'* Today the principle seems self-evident: Only research with the ability to be reproduced allows for confirmation of the claims and conclusions of the researchers. Research without this ability can be attributed to either (1) mistakes, errors, random chance, (2) malicious fraud, or (3) that the results are valid only under certain conditions. The first two imply that the supposed findings are, in fact, wrong, whereas the latter yield new knowledge about the conditions under which it holds true [20].

However, there are many ways to reproduce experiments. In machine learning research, Drummond [16] argues to separate reproducibility from *'its poor cousin'* of replicability. In essence, *reproducibility* refers to the desirable trait of simply being able to reproduce the conclusions of a study. One can easily imagine results and conclusions reproduced using an entirely different experimental design, technology, and statistical methodology. In contrast, *replicability* refers

Table 1: Suggested terminology and flavors of reproducibility.

Term	Explanation
Repetition	Rerun exactly what others have done, reusing their experimental data.
Replication	Redo the exact experiment and data analysis of what others have done.
Variation	Redo or replicate what others have done, but with slight modifications in experimental setup or data analysis.
Reproduction	Recreate the spirit of what others have done.
Corroboration	Arrive at the same conclusions using an entirely different experimental design and/or data analysis.
Validation	Confirm specific results from a previous analysis in an independent experiment.

to the ability to exactly replicate the results and output of a given study, i.e. reproduce exactly the results given the experimental outcome. However, we must commit the yet more detailed terminology heavily inspired by Feitelson [20] presented in Table 1. The essence of some of the terms have been modified slightly to suit our needs along with the addition of *validation*. Note, the terms are used inconsistently in the literature. Within the *papers* of this thesis, only *reproducibility* in its general sense is used and the appropriate specific term can be substituted depending on the context.

It is also worth stressing the meaning of *internal* and *external* which can be prefixed to the terms Table 1. Internal replication refers to replications by the same researchers, whereas external necessitates independent researchers to perform the replication.

3.2 A continuum of reproducibility

There is, of course, no exact well-defined criteria of repeatability, replicability, and reproducibility. Though distinct, the terms may be quite overlapping in many cases. Likewise, each term is not a dichotomy. A given experiment and its analysis is not simply repeatable or not. Instead a continuous scale can be imagined which indicate the degree to which they submit to the term. For repeatability of the analysis, the scale indicates how easy when given raw data, the results, output, and hence scientific conclusions are repeated. At one end of this scale, the results and output are repeated automatically by the press of a button. At the other end are poorly documented and disconnected workflows distributed on many collaborating researchers contributing via different analysis platforms. Examples are graphical user interfaces, where the exact actions are hard to document, and thus errors are very likely to go unnoticed. The

latter offers virtually no possibility on retracing and replicating mistakes and errors. Baggerly and Coombes [2] discussed the frighteningly real consequences of opaque, sloppy, and non-replicable research of Potti et al. [44] and related papers and their time-consuming use of 'forensic bioinformatics' to infer the methods used in the poorly documented research with unhelpful researchers. Potti et al. [44] used microarrays and cell line drug sensitivity data to predict the sensitivity of patients to various chemotherapeutics that, if true, presented a significant step towards personalized medicine. The publication ultimately instigated clinical trials potentially putting patients at risk However, following Coombes et al. [9] the case evolved into the infamous Duke University scandal, and the articles have since been retracted and the clinical trials stopped. Even though the case also included fraud, which is arguably rare [16, 23], the poorly documented research unfortunately is not a one-off [3, 27, 56].

For replicability, the scale indicates how well described the experimental setup, statistical methods, etc., are. Exact replicability may often be infeasible or impossible as patients can die, emigrate, or otherwise fail to participate in a replication study. On the other hand, using exactly the same patients are hardly implied when talking about replication studies; this might more naturally fall under the repetition term.

As statistical methods analyzing high-dimensional data, biological or otherwise, are often quite involved requiring many steps of data preprocessing, analyses, and output generation, replicability in the form of a completely documented workflow is very desirable. The many statistical and probabilistic pitfalls only serve as further arguments as to why replicability is needed on the road to reproducibility. Non-repeatable or non-replicable mistakes and errors are many times worse than repeatable and replicable ones.

Repeatability and replicability

Sonnenburg et al. [54] advocate for complete open source as the means for reproducibility while Drummond [16], however, concludes *'that replicability is a poor substitute for scientific reproducibility. There may be other good reasons for the collecting of software and scripts that are the basis of the experimental results published in papers but scientific reproducibility is not one.'*

There is no doubt a hierarchy of importance of the terms. While reproducibility is the main property, replicability is not as futile a goal as Drummond [16] makes it seem. It is desirable in itself and worth the extra effort as exemplified above. Indeed, replicability is no substitute for reproducibility. Replicability does not imply reproducibility, nor does reproducibility imply replicability. Replicability is necessary for others to ascertain the precise details of the scientific work, which is often not of particular interest in itself. It is, however, of interest *precisely when others cannot reproduce* the scientific work. Particularly, it allows elucidating *why* the experimental results do not reproduce and, better yet, determine under which conditions the results do hold true. Drummond [16] argues the only seeming value of replicability is for

prevention of scientific fraud, which is a minor one due to the rareness of it.

There are indeed plenty of other good reasons for collecting software and scripts. The improved replicability from computer code greatly increases transparency and provides *exact* documentation. Sonnenburg et al. [54] argues that it also implies faster scientific progress, reduced costs, and quicker detection and correction of scientific mistakes and software bugs.

3.3 Reasons for poor reproducibility

The precise reason(s) for the discrepancies in scientific results are poorly understood. Among the numerous proposed explanations are: (1) The number of explanatory variables p always far exceed the sample size n, i.e. $p \gg n$, which renders all practical sample sizes comparatively small. Very large sample sizes are infeasible for most research groups due to heavy costs, time constraints, computational costs, or lack of biological material. (2) Relatively few signals are often truly present in the underlying stochastic process from which the data is created. (3) The data are subject to substantial variability by a vast number of biological and technical factors. (4) Crucial genes near the significance threshold may often be left out. (5) Experiments differ on an important, possibly confounding, factor, e.g. age of cohort or the technology used. (6) Differing statistical methodologies. While the latter may decrease the apparent level of reproducibility, it should not destroy it completely. Surely, true biological findings should be reasonably robust to different statistical methodologies and experimental setup. Indeed this can also be said for the other reasons listed.

Lastly, there seems to be a general tendency in high-throughput biological experiments for money to be invested in newer profiling technologies, deeper sequencing, broader searches, or more subgroups, rather than simply obtaining a larger *sample size*. This tendency, also noted by Robinson and Smyth [47, p. 2881], further amplifies the problems of reproducibility.

At the heart of the high-dimensional $p \gg n$ problems is also the risk of overfitting. By the sheer number of observations, the variable of interest can *always* be perfectly explained by *some* combination of them—even by random noise. This is, essentially, the so-called overfitting where the statistical model incorporates random noise rather than the true underlying relationship. Hence, high-throughput screenings and searches are prone to produce false scientific findings, since low signal-to-noise ratios are intrinsic to such experiments.

3.4 Fishing for significance and publication bias

Other sources of irreproducibility are publication bias and researcher bias, consciously or unconsciously, for statistically significant results. The former is a well-known problem driven by the tendency of scientific journals to accept positive (statistically significant) results with a higher likelihood [24, 57]. This is coupled with reluctance to submit negative results by the scientists. It is also known as the *file drawer problem* as many 'negative' results are often consid-

ered uninteresting and unpublishable, and thus they end their lives in dusty file drawers. The problem may easily be trivialized, but it is hugely important due to the observation that the majority of all results likely are negative [12]. Hence, the systematic omission of negative results potentially greatly increases the likelihood of false positive findings in published results. Furthermore, publishing negative results is also important, as they indicate what does not work. A number of journals, listed in da Silva [12], exclusively devoted to negative results have been made to combat the file drawer problem.

The researcher bias has also been more descriptively called *fishing for significance* or even *p-value hacking* [24]. The problem is, that researchers often report results from an optimization process searching for methods (or variants of the method) that work better, yield results that are more significant, lower error rates, or whatever measure that suit the application. This process can be entirely subconscious. If it is conscious it can be benign; malicious intent is not necessarily implied. After all, there is nearly always an element of trial-and-error in developing good statistical methods [6]. The problem is that it is, in essence, the multiple hypothesis testing problem in disguise—*'If you torture the data long enough, it will confess'* as famously put by Ronald H. Coase. Boulesteix [6] argues that the difficulty in publishing negative results might actually encourage and amplify this p-hacking tendency.

Boulesteix [6] gives a good summarization of these issues and concludes that publishing scripts for replicability and reproducibility contributes to more transparent research and rapid unbiased validation of results. For example, parameters that were tuned, consciously or subconsciously, are more easily identified. It does not ensure that authors have not overfitted or p-hacked; it ensures that these issues are more easily tracked down.

3.5 Tools for replicability and data integration

In the statistical programming language R [45], the concept of easy replicability is not novel. **Sweave** [33], later extended by **knitr** [69], is based on the literate programming paradigm that integrates R with LaTeX, HTML, mark-down and other mark-up languages. This serves as easily replicable results incorporated into the manuscripts, reports, or books. It also has the added benefit of an instantaneously updated manuscript if e.g. mistakes and errors in the raw data, method, or software are corrected.

There are many other tools available towards the goal of exact and easy replication of results. The R-package **packrat** aids in managing the version of the used software packages and their dependencies [62]. Online code repositories and version control hosting sites such as `bitbucket.org` and `github.com` provide excellent means for moving toward complete open sourced science and easier collaboration with other researchers.

Online data repositories also increase the replicability of studies allowing other researchers to download the publicly available data. One example is the Gene Expression Omnibus (GEO) from the National Center for Biotechno-

logy Information (NCBI), which currently holds nearly 4,000 public datasets amounting to more than 1.5 million samples. The R-package **GEOquery** [13] provides a programmatic interface to GEO.

Considerable effort has been invested to make the papers of this thesis, the replication of the results, and the raw data easily and publicly available to the biological and statistical community via some of the tools above. To this end, the package **DLBCLdata** is an attempt to provide preprocessing of DLBCL gene expression datasets in an easy, automated, and reproducible manner.

3.6 Data integration and meta-analysis

Meta-analysis and data integration seek to aggregate information across multiple studies to achieve a higher sensitivity and specificity. As such, both hold important aspects of reproducibility.

Traditional meta-analysis works on combining summary quantities such as test statistics, confidence intervals, p-values, etc. The GMCM of Paper I can alternatively be thought of as a novel method for aggregating or combining test statistics. A variation of meta-analysis is pooling the raw data of the studies in a sometimes so-called *mega*-analysis [See e.g. 46]. Papers III and IV presents methods for mega-analysis in network analysis.

Meta-analysis is not entirely uncontroversial [19]. In all meta-analyses, the publication bias is a well-known issue. However, the comparability of the studies should also be considered, as to not compare *apples and oranges*. Using the raw data of large-scale microarray studies publication bias is likely less of an issue and the p-hacking issue seems to vanish.

3.7 Validation experiments

The poor reproducibility of high-dimensional experiments has manifested itself as a widespread acknowledgement of the need of validation experiments. Already in 1999, near the introduction of microarrays, Brown and Botstein [7] stressed that microarray analyses are *exploratory*, not driven by hypotheses, and should therefore be as independent of the model as possible. In this view, the statistical screenings are viewed as *hypothesis generating*; the statistical outcomes are nothing more than a prioritized list of promising candidates. Hence, the most interesting suggestions from the screenings needs to be validated in independent experiments using the traditional gold-standard methods.

The term *hypothesis generating* for high-dimensional screens is particularly useful in avoiding many reproducibility issues. The name directly implies the need for validation as only *hypotheses* are created from high-dimensional screenings.

Paper V is central to this subject and provides a correction to the potentially overly optimistic statistical analysis of so-called qPCR experiments.

4 Overview of the thesis

In the following part of this thesis the main papers are presented. The papers have been reformatted and minor grammar and spelling corrections have been made. An introductory preface have been added to each paper relating the material presented above. Following the papers, a preface to the statistical software packages written is also given.

Paper I & II Paper I discusses two classes of so-called Gaussian mixture copula models (GMCM). The first GMCM was proposed by Li et al. [35] and a generalized GMCM was independently proposed by Tewari et al. [60], dubbed the special and general model, respectively. GMCMs are attractive as they are highly flexible. The former special model was proposed for quantifying the degree to which genes of two or more gene-lists agree. In this sense, the special model is capable of aggregating evidence across multiple studies of differential expression. However, a number of issues were identified in the rather slow implementation in the R-package **idr** by Li et al. [35]. Paper I and Package I serve as an introduction to GMCMs, solves some of the identified issues, and provides very fast estimation procedures. The special model was also applied to the data of Paper II as a secondary analysis example.

The general model of Tewari et al. [60] was also incorporated and suggested for general purpose unsupervised clustering. One point of expansion of the general GMCM is to apply the model in the closely related discriminant analysis also discussed and applied in Paper III.

Paper II is an applied paper concerned with differential expression of genes between cryopreserved (frozen) and non-cryopreserved (fresh) samples of white blood cells. Cryopreservation is a very convenient tool for biological researchers to store tissue or cells and thus postpone analysis or use. It is a fundamental part of bio-banks that store biological samples and tissue. Naturally, a premise of using cryopreserved samples in research is that the results would be largely unchanged had fresh tissue been used instead. Or, results will reproduce when using cryopreserved tissue compared to that of fresh tissue. This paper investigates the validity of that premise and concludes that few genes are changed due to the cryopreservation.

Paper III & IV The papers III and IV in this thesis describe statistical models usable in the attempt to elucidate genetic networks described above. They can both be used in Gaussian graphical modelling or for identifying correlation networks across many datasets. As network estimation is known to be quite variable, these papers attempt to integrate the information of many datasets to stabilize the estimation and arrive at more reproducible results.

Paper V Paper V describes an easily corrected omission in the standard statistical analysis of so-called qPCR experiments; the gold standard often used to validate findings from high-throughput experiments. The current statistical methods ignore a potentially important source of variation that leads to overly optimistic analyses. Besides providing ways to incorporate the omitted variance, the paper also provides a unified statistical framework for analyzing such

qPCR data.

Package I The R-package **GMCM** is the accompanying software to Paper I.

Package II The R-package **DLBCLdata** is used in papers III and IV. It automates the otherwise very cumbersome process of manually downloading datasets and scripting the preprocessing hereof. This results in completely replicable analyses. The package 'features' 12 large-scale gene expression datasets of diffuse large B-cell lymphoma (DLBCL) cancer but should work for most GEO datasets using Affymetrix platforms.

Package III The R-package **rags2ridges** is the accompanying implementation of [63] and now also Paper I. Much of the base-code in **rags2ridges** were rewritten in C++ to accommodate the more computationally demanding *fused* variant of ridge precision estimation.

References

[1] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, 2000.

[2] K. A. Baggerly and K. R. Coombes. Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *The Annals of Applied Statistics*, pages 1309–1334, 2009.

[3] K. A. Baggerly and K. R. Coombes. What information should be required to support clinical "omics" publications? *Clinical chemistry*, 57(5):688–690, 2011.

[4] C. G. Begley and L. M. Ellis. Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533, Mar. 2012. ISSN 0028-0836. URL http://dx.doi.org/10.1038/483531a.

[5] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(1):289–300, 1995.

[6] A.-L. Boulesteix. Over-optimism in bioinformatics research. *Bioinformatics*, 26(3):437–439, 2010. Letter to the editor.

[7] P. O. Brown and D. Botstein. Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, 21:33–37, 1999.

[8] M. A. Care, S. Barrans, L. Worrillow, A. Jack, D. R. Westhead, and R. M. Tooze. A microarray platform-independent classification tool for cell of origin class allows comparative analysis of gene expression in diffuse large B-cell lymphoma. *PLoS One*, 8(2):e55895, 2013.

[9] K. R. Coombes, J. Wang, and K. A. Baggerly. Microarrays: retracing steps. *Nature medicine*, 13(11):1276–1277, 2007.

[10] F. Crick et al. Central dogma of molecular biology. *Nature*, 227(5258): 561–563, 1970.

[11] F. H. Crick. On protein synthesis. In *Symposia of the Society for Experimental Biology*, volume 12, page 138, 1958.

[12] J. A. T. da Silva. Negative results: Negative perceptions limit their potential for increasing reproducibility. *Journal of negative results in biomedicine*, 14(1):12, 2015.

[13] S. Davis and P. Meltzer. **GEOquery**: A bridge between the gene expression omnibus (GEO) and bioconductor. *Bioinformatics*, 14:1846–1847, 2007.

[14] R. de Sousa Abreu, L. O. Penalva, E. M. Marcotte, and C. Vogel. Global signatures of protein and mRNA expression levels. *Molecular BioSystems*, 5(12):1512–1526, 2009.

[15] A. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972.

[16] C. Drummond. *Replicability is not Reproducibility: Nor is it Good Science*, 2009. URL http://cogprints.org/7691/. Proc. of the Evaluation Methods for Machine Learning Workshop at the 26th ICML, Montreal, Canada.

[17] S. Dudoit, J. Shaffer, and J. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, pages 71–103, 2003.

[18] L. Ein-Dor, O. Zuk, and E. Domany. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 103 (15):5923–8, Apr. 2006. ISSN 0027-8424. doi: 10.1073/pnas.0601231103.

[19] A. R. Feinstein. Meta-analysis: Statistical alchemy for the 21st century. *Journal of clinical epidemiology*, 48(1):71–79, 1995.

[20] D. G. Feitelson. From repeatability to reproducibility and corroboration. *ACM SIGOPS Operating Systems Review*, 49(1):3–11, 2015.

[21] R. A. Fisher. *The Design of Experiments*. Hafner Publishing Company, Inc, 8th edition, 1966 edition, 1935.

[22] W. W. Gibbs. The unseen genome: Gems among the junk. *Scientific American*, 289(5):46–53, 2003.

[23] W. Gunn. Reproducibility: Fraud is not the big problem. *Nature*, 505 (7484):483, Jan. 2014. ISSN 0028-0836. URL http://dx.doi.org/10.1038/505483b10.1038/505483b.

[24] M. L. Head, L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions. The extent and consequences of p-hacking in science. *PLoS Biology*, 13 (3):e1002106, 2015.

[25] S. Henikoff. Beyond the central dogma. *Bioinformatics*, 18(2):223–225, 2002.

[26] S. Horvath. *Weighted Network Analysis: Applications in Genomics and Systems Biology*. Springer, 2011.

[27] D. Ince. The duke university scandal—what can be done? *Significance*, 8 (3):113–115, 2011.

[28] International Lymphoma Study Group. A clinical evaluation of the international lymphoma study group classification of non-hodgkin's lymphoma. *Blood*, 89(11):3909–3918, June 1997. The Non-Hodgkin's Lymphoma Classification Project.

[29] J. P. Ioannidis, E. E. Ntzani, T. a. Trikalinos, and D. G. Contopoulos-Ioannidis. Replication validity of genetic association studies. *Nature genetics*, 29(3):306–9, Nov. 2001. ISSN 1061-4036. doi: 10.1038/ng749.

[30] J. P. A. Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, Aug. 2005. ISSN 1549-1676. doi: 10.1371/journal.pmed.0020124.

[31] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[32] S. L. Lauritzen. *Graphical models*. Clarendon Press, Oxford, 1996.

[33] F. Leisch. `Sweave`: Dynamic generation of statistical reports using literate data analysis. In W. Härdle and B. Rönz, editors, *Compstat 2002 — Proceedings in Computational Statistics*, pages 575–580. Physica Verlag, Heidelberg, 2002. ISBN 3-7908-1517-9.

[34] G. Lenz, G. Wright, S. Dave, W. Xiao, J. Powell, H. Zhao, W. Xu, B. Tan, N. Goldschmidt, J. Iqbal, and Others. Stromal gene signatures in large-B-cell lymphomas. *New England Journal of Medicine*, 359(22):2313–2323, 2008.

[35] Q. Li, J. B. J. Brown, H. Huang, and P. J. Bickel. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics*, 5 (3):1752–1779, 2011. ISSN 1932-6157. doi: 10.1214/11-AOAS466.

[36] W. Li, C.-C. Liu, T. Zhang, H. Li, M. S. Waterman, and X. J. Zhou. Integrative analysis of many weighted co-expression networks using tensor computation. *PLoS Computational Biology*, 7(6):e1001106, 2011.

[37] T. Maier, M. Güell, and L. Serrano. Correlation of mrna and protein in complex biological samples. *FEBS letters*, 583(24):3966–3973, 2009.

[38] F. Markowetz and R. Spang. Inferring cellular networks—a review. *BMC Bioinformatics*, 8(Suppl 6):S5, 2007.

[39] J. S. Mattick. Challenging the dogma: The hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays*, 25(10):930–939, 2003.

[40] T. Mehta, M. Tanik, and D. B. Allison. Towards sound epistemological foundations of statistical methods for high-dimensional biology. *Nature genetics*, 36(9):943–947, 2004.

[41] G. S. Nowakowski, B. LaPlant, W. R. Macon, C. B. Reeder, J. M. Foran, G. D. Nelson, C. A. Thompson, C. E. Rivera, D. J. Inwards, I. N. Micallef, P. B. Johnston, L. F. Porrata, S. M. Ansell, R. D. Gascoyne, T. M. Habermann, and T. E. Witzig. Lenalidomide combined with R-CHOP overcomes negative prognostic impact of non-germinal center B-cell phenotype in newly diagnosed diffuse large B-cell lymphoma: A phase II study. *Journal of Clinical Oncology*, 33(3):251–257, 2015.

[42] H. B. Pedersen. *Kræftens Biologi*. Forlaget Systime, 2001. ISBN 87 7783-663-4.

[43] K. Popper. *The Logic of Scientific Discovery*. Routledge, 1934. ISBN 0-203-99462-0. 2005 Routledge Edition.

[44] A. Potti, H. K. Dressman, A. Bild, R. F. Riedel, G. Chan, R. Sayer, J. Cragun, H. Cottrill, M. J. Kelley, R. Petersen, et al. Genomic signatures to guide the use of chemotherapeutics. *Nature medicine*, 12(11):1294–1300, 2006.

[45] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL http://www.R-project.org/. ISBN 3-900051-07-0.

[46] S. Ripke, N. R. Wray, C. M. Lewis, S. P. Hamilton, M. M. Weissman, G. Breen, E. M. Byrne, D. H. Blackwood, D. I. Boomsma, S. Cichon, et al. A mega-analysis of genome-wide association studies for major depressive disorder. *Molecular psychiatry*, 18(4):497–511, 2013.

[47] M. D. Robinson and G. K. Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, 2007.

[48] A. Rosenwald, G. Wright, W. C. Chan, J. M. Connors, E. Campo, R. I. Fisher, R. D. Gascoyne, H. K. Muller-Hermelink, E. B. Smeland, J. M. Giltnane, E. M. Hurt, H. Zhao, L. Averett, L. Yang, W. H. Wilson, E. S. Jaffe, R. Simon, R. D. Klausner, J. Powell, P. L. Duffey, D. L. Longo, T. C. Greiner, D. D. Weisenburger, W. G. Sanger, B. J. Dave, J. C. Lynch, J. Vose, J. O. Armitage, E. Montserrat, A. López-Guillermo, T. M. Grogan, T. P. Miller, M. LeBlanc, G. Ott, S. Kvaloy, J. Delabie, H. Holte, P. Krajci, T. Stokke, and L. M. Staudt. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *The New England Journal of Medicine*, 346(25):1937–47, June 2002. ISSN 1533-4406. doi: 10.1056/NEJMoa012914.

[49] J. Ruan, P. Martin, R. R. Furman, S. M. Lee, K. Cheung, J. M. Vose, A. LaCasce, J. Morrison, R. Elstrom, S. Ely, A. Chadburn, E. Cesarman, M. Coleman, and J. P. Leonard. Bortezomib plus CHOP-rituximab for previously untreated diffuse large B-cell lymphoma and mantle cell lymphoma. *Journal of Clinical Oncology*, 29(6):690–697, 2011.

[50] B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342, 2011.

[51] S. Shapin and S. Schaffer. *Leviathan and the air-pump: Hobbes, Boyle, and the experimental life*. Princeton: Princeton University Press, 1985.

[52] R. M. Simon, E. L. Korn, L. M. McShane, M. D. Radmacher, G. W. Wright, and Y. Zhao. *Design and Analysis of DNA Microarray Investigations*. Springer-Verlag, 2003. ISBN 0-387-00135-2.

[53] G. K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1):1–25, 2004.

[54] S. Sonnenburg, M. L. Braun, C. S. Ong, S. Bengio, L. Bottou, G. Holmes, Y. LeCun, K.-R. Müller, F. Pereira, C. E. Rasmussen, et al. The need for open source software in machine learning. *Journal of Machine Learning Research*, 8:2443–2466, 2007.

[55] T. P. Speed, editor. *Statistical Analysis of Gene Expression Microarray Data*. Chapman & Hall/CRC, 2003. ISBN 1-58788-327-8.

[56] R. G. Steen. Retractions in the medical literature: How many patients are put at risk by flawed research? *Journal of Medical Ethics*, pages jme–2011, 2011.

[57] T. D. Sterling. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American statistical association*, 54(285):30–34, 1959.

[58] S. H. Swerdllow, E. Campo, N. L. Harris, E. S. Jaffe, S. A. Pileri, S. H, J. Thiele, and J. W. Vardiman. *WHO classification of tumours of haematopoietic and lymphoid tissues*. Lyon: IARC Press, 2008, 2008.

[59] P. Tan, T. Downey, and E. S. Jr. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Research*, 31 (19):5676–5684, Oct. 2003. ISSN 1362-4962. doi: 10.1093/nar/gkg763.

[60] A. Tewari, M. J. Giering, and A. Raghunathan. Parametric characterization of multimodal distributions with non-gaussian modes. *ICDM 2011 conference*, pages 286–292, Dec. 2011. doi: 10.1109/ICDMW.2011.135.

[61] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

[62] K. Ushey, J. McPherson, J. Cheng, and J. Allaire. *packrat: A Dependency Management System for Projects and their R Package Dependencies*, 2015. URL `http://CRAN.R-project.org/package=packrat`. R package version 0.4.3.

[63] W. N. van Wieringen and C. F. W. Peeters. Ridge estimation of inverse covariance matrices from high-dimensional data. Submitted to *Computational Statistics & Data Analysis, arXiv:1403.0904v3*, 2015.

[64] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.

[65] C. Visco, Y. Li, Z. Y. Xu-Monette, R. N. Miranda, T. M. Green, A. Tzankov, W. Wen, W.-m. Liu, B. S. Kahl, E. S. G. D'Amore, S. Montes-Moreno, K. Dybkæ r, A. Chiu, W. Tam, A. Orazi, Y. Zu, G. Bhagat, J. N. Winter, H.-Y. Wang, S. O'Neill, C. H. Dunphy, E. D. Hsi, X. F. Zhao, R. S. Go, W. W. L. Choi, F. Zhou, M. Czader, J. Tong, X. Zhao, J. H. van Krieken, Q. Huang, W. Ai, J. Etzell, M. Ponzoni, a. J. M. Ferreri, M. a. Piris, M. B. Mø ller, C. E. Bueso-Ramos, L. J. Medeiros, L. Wu, and K. H. Young. Comprehensive gene expression profiling and immunohistochemical studies support application of immunophenotypic algorithm for molecular subtype classification in diffuse large B-cell lymphoma: a report from the international DLBCL rituximab-CHOP consortium. *Leukemia*, 26(9):2103–13, Sept. 2012. ISSN 1476-5551. doi: 10.1038/leu.2012.83.

[66] J. D. Watson et al. *Molecular Biology of the Gene*. New York: WA Benjamin, Inc., 2nd edn edition, 1970.

[67] G. Wright, B. Tan, A. Rosenwald, E. Hurt, A. Wiestner, and L. Staudt. A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *Proceedings of the National Academy of Sciences of the United States of America*, 100(17):9991, Aug. 2003. ISSN 0027-8424. doi: 10.1073/pnas.1732008100.

[68] Z. Wu. A review of statistical methods for preprocessing oligonucleotide microarrays. *Statistical methods in medical research*, 18(6):533–541, 2009.

[69] Y. Xie. *Dynamic Documents with* R *and* **knitr**. CRC Press, 2013. ISBN 9781482203530.

[70] M. Zhang, C. Yao, Z. Guo, J. Zou, L. Zhang, H. Xiao, D. Wang, D. Yang, X. Gong, J. Zhu, Y. Li, and X. Li. Apparently Low Reproducibility of True Differential Expression Discoveries in Microarray Studies. *Bioinformatics*, 24(18):2057–63, Sept. 2008. ISSN 1367-4811. doi: 10.1093/bioinformatics/btn365.

Part B

Papers

Paper I

Unsupervised Clustering and Meta-Analysis using Gaussian Mixture Copula Models

Anders Ellern Bilgrau, Poul Svante Eriksen, Jakob Gulddahl Rasmussen, Hans Erik Johnsen, Karen Dybkær, and Martin Bøgsted

Accepted for publication in the
Journal of Statistical Software, 2015.

Preface: This paper generalizes work by Li et al. [21] along the lines of Tewari et al. [33]. The paper discusses the class of so-called Gaussian mixture copula models (GMCM), how to estimate the parameters hereof, and the software implementation. A GMCM is a multivariate Gaussian mixture model that is invariant to all monotone marginal transformations. This property comes from the fact that the model only cares about the marginal *rankings* of the observations—and it makes the model quite flexible, but also harder to fit.

The Li et al. [21] special case attempts to quantify the level of reproducibility for each gene in ranked lists of genes from different experiments. The quantity is an *irreproducibility discovery rate* which can be thought of as an integrated score used to infer genes as either reproducible or irreproducible. The irreproducibility discovery rates are essentially local false discovery rates also used in Paper IV. The generalized model of Tewari et al. [33] uses analogous quantities for determining the classes in unsupervised clustering.

The GMCM may be adopted for discriminant analysis, also used in Paper III, as one point of extension. Further work could be determination of the number of clusters from data or a regularized GMCM, similar to Paper IV or otherwise.

The accompanying R-package **GMCM**, also presented in Package I, is freely available on CRAN (the comprehensive R archive network) at

`http://cran.r-project.org/package=GMCM`

while the development branch is found at

`https://github.com/AEBilgrau/GMCM`.

The package is open-source and this paper is also found as a package vignette by following the former link.

 The layout has been revised.

Unsupervised Clustering and Meta-Analysis using Gaussian Mixture Copula Models

Abstract

Methods for unsupervised clustering is an important part of the statistical toolbox in numerous scientific disciplines. Tewari et al. [33] proposed to use so-called Gaussian Mixture Copula Models (GMCM) for general unsupervised clustering. Li et al. [21] independently discussed a special case of these GMCMs as a novel approach to meta-analysis in high-dimensional settings. GMCMs have attractive properties which make them highly flexible and therefore interesting alternatives to well-established methods. However, parameter estimation is hard because of intrinsic identifiability issues and intractable likelihood functions. Both aforementioned papers discuss similar expectation-maximization-like (EM) algorithms as their pseudo maximum likelihood estimation procedure. We present and discuss an improved implementation in R of both classes of GMCMs along with various alternative optimization routines to the EM algorithm. The software is freely available through the accompanying R package **GMCM**. The implementation is fast, general, and optimized for very large numbers of observations. We demonstrate the use of **GMCM** through different applications.

1 Introduction

Unsupervised cluster analysis is an important discipline in many fields of science and engineering for detection of clusters of data with similar properties. Gaussian mixture models (GMM) is perhaps the most widely used method for unsupervised clustering of continuous data. However, the assumption of jointly normally distributed clusters in GMMs is often violated. Tewari et al. [33] presented the semi-parametric class of Gaussian mixture copula models (GMCM) for general unsupervised clustering and highlighted them as a flexible alternative to GMMs when obvious non-normally distributed clusters are present. The attractiveness of the GMCMs is predominantly due to an invariance under all monotone increasing marginal transformations of the variables. This scale invariance of the variables stems from the rank-based nature of copula models and make the GMCMs highly versatile.

The GMCMs have found some success in applications after Li et al. [21] independently proposed using a special-case for a non-standard meta-analysis methodology named reproducibility analysis. Their method have been adopted by the ENCODE project [4, 34] and applied on ChIP-sequencing data. The meta-analysis approach with GMCMs works by clustering genes or features that agree on statistical evidence and those that do not. In other words, the features are clustered into a reproducible and an irreproducible group. The

flexibility of the GMCMs make them suitable for meta-analysis of multiple similar experiments.

The work of Li et al. [21] is especially important in genomics as both data and results are subject to substantial variability due to limited samples sizes, high dimensional feature spaces, dependence between genes, and confounding technological factors. This high variability have brought into question the reliability and reproducibility of many genomic results [13, 17, 32]. Others, however, argue that the lack of reproducibility is only superficial [37]. Together with a rapid evolution of many different high-throughput technologies and vast online repositories of publicly available data, this motivates the need for a robust and flexible meta-analysis toolbox, which can evaluate or aggregate results of multiple experiments even across confounding factors such as differing technologies.

The high flexibility of the GMCMs comes at a cost, however. The likelihood is difficult to evaluate and maximize, partly because of intrinsic identifiability problems as we describe in detail later. We have solved some of the issues and implemented them in the package **GMCM** for R [27].

Although copula theory is an elegant way of approaching rank-based methods, we present the GMCMs in a more traditional fashion. We refer to the general model of Tewari et al. [33] simply as the *general model* or *general GMCM* and the special case model of Li et al. [21] is referred to as the *special model* or *special GMCM*.

In the following, we present the general GMCM followed by the special case and the derivation of the likelihood function. Subsequently, the key features of the **GMCM** software package are presented and compared to the **idr** package. The technical details of the problematic maximization of the likelihood are then discussed. Finally, our package is evaluated by different applications before concluding with a discussion of GMCMs.

This document was prepared and generated using **knitr** [36], a dynamic report generation tool inspired by Sweave [19], and the R-packages **Hmisc** [15] and **RColorBrewer** [24]. The simulation study was carried out using parallel computing with **doMC** [28] and **foreach** [29].

2 Gaussian mixture copula models

2.1 The general GMCM for unsupervised clustering

We consider a large $p \times d$ matrix $[x_{gk}]$ of observed values where the rows are to be clustered into m groups. The general GMCM assumes an m-component Gaussian mixture model (GMM) as a latent process, $\mathbf{Z} = (Z_1, \dots, Z_d)^\top$, with the following distribution

$$\text{GMM:} \qquad \begin{cases} H \sim \text{Categorical}(\alpha_1, \dots, \alpha_m) \\ \mathbf{Z}|H = h \sim \mathcal{N}_d(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h) \end{cases} \tag{1}$$

where $H \in \{1, 2, \dots, m\}$ corresponds to the class and $\alpha_1, \dots, \alpha_m$ are the mixture proportions satisfying $\sum_{h=1}^m \alpha_h = 1$. Thus, the latent GMM is parameterized by

$$\boldsymbol{\theta} = (\alpha_1, \dots, \alpha_m, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_m).$$

We denote the joint and k'th marginal cumulative distribution functions (cdf) of the GMM by

$$\Gamma(\mathbf{z}; \boldsymbol{\theta}) = \sum_{h=1}^m \alpha_h \Phi(\mathbf{z}; \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h) \quad \text{and} \quad \Gamma_k(z; \boldsymbol{\theta}) = \sum_{h=1}^m \alpha_h \Phi_k(z; \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h),$$

respectively, where Φ and Φ_k are the joint and k'th marginal cdfs of the multivariate normal distributions, respectively. Analogous equations hold for the joint and marginal probability density functions (pdf) which we denote by lower-case γ and γ_k.

Let $\mathbf{X} = (X_1, \dots, X_d)^\top$ be an observation with *known* marginal cumulative distribution functions $F_1, \dots, F_d$ and assume the relationship

$$X_k = F_k^{-1}\big(\Gamma_k(Z_k; \boldsymbol{\theta})\big), \quad \forall k \in \{1, \dots, d\} \tag{2}$$

between the observed and the latent variables. By Equation 2 and the probability integral transform the vector $\mathbf{U} = (U_1, \dots, U_d)^\top$ where $U_k = \Gamma_k(Z_k) = F_k(X_k)$ have uniformly distributed marginals.

When $F_1, \dots, F_d$ are known we can derive an expression for the likelihood of this model. For later use we simplify the notation by introducing the vector functions $\Gamma_\circ : \mathbb{R}^d \times \Theta \to \mathbb{R}^d$ and $F_\circ : \mathbb{R}^d \to \mathbb{R}^d$ defined by

$$\Gamma_\circ(\mathbf{Z}; \boldsymbol{\theta}) = \big(\Gamma_1(Z_1; \boldsymbol{\theta}), \dots, \Gamma_d(Z_d; \boldsymbol{\theta})\big)^\top \quad \text{and} \quad F_\circ(\mathbf{X}) = \big(F_1(X_1), \dots, F_d(X_d)\big)^\top,$$

where Θ is the parameter space. The vector function $\Gamma_\circ$ applies the k'th marginal transformation Γ_k on the k'th entry of the observation and similarly does $F_\circ$. Again by the probability integral transform, $\mathbf{Z}$ is transformed by $\Gamma_\circ$ into the marginally uniformly distributed random vector $\mathbf{U}$ with cdf

$$C(\mathbf{u}; \boldsymbol{\theta}) = \Gamma\big(\Gamma_\circ^{-1}(\mathbf{u}; \boldsymbol{\theta}); \boldsymbol{\theta}\big).$$

The pdf c of $\mathbf{U}$ is computed by the change of variables theorem or by differentiation of C using the multivariable chain rule. If we abbreviate notationally by not explicitly stating dependence on parameters $\boldsymbol{\theta}$, the pdf is given by

$$c(\mathbf{u};\boldsymbol{\theta}) = \gamma\big(\Gamma_\circ^{-1}(\mathbf{u})\big)\left|J_{\Gamma_\circ^{-1}}(\mathbf{u})\right| = \frac{\gamma\big(\Gamma_\circ^{-1}(\mathbf{u})\big)}{\prod_{k=1}^{d}\gamma_k\big(\Gamma_k^{-1}(u_k)\big)} \tag{3}$$

since the Jacobian matrix $J_{\Gamma_\circ^{-1}}(\mathbf{u})$ is diagonal. The cdf C and pdf c are the so-called *copula* and *copula density* of the GMM model, respectively [23]. Hence $\mathbf{U}$ is distributed according to the Gaussian mixture copula density c, and the observation $\mathbf{X}$ is some marginal transformation of $\mathbf{U}$. The model is thus completely specified by

$$\text{GMCM:} \quad \begin{cases} H \sim \text{Categorical}(\alpha_1,\ldots,\alpha_m) \\ \mathbf{Z}|H=h \sim \mathcal{N}_d(\boldsymbol{\mu}_h,\boldsymbol{\Sigma}_h) \\ \mathbf{U} = \Gamma_\circ(\mathbf{Z};\boldsymbol{\theta}) \\ \mathbf{X} = F_\circ^{-1}(\mathbf{U}). \end{cases} \tag{4}$$

From this, we see the GMCM operates on three levels, a latent level $\mathbf{Z}$, a copula level $\mathbf{U}$, and an observed level $\mathbf{X}$. Figure 1 (A-C) illustrates the three levels of a 2-dimensional 3-component GMCM. Here, $F_\circ$ and $F_\circ^{-1}$ maps panel A to B and B to A, respectively. Likewise, $\Gamma_\circ$ defines the mappings between panels C and B.

To assess the class of an observation, Tewari et al. [33] proposed using

$$\kappa_h = P(H = h \mid \mathbf{u},\boldsymbol{\theta}), \tag{5}$$

which is the a posteriori probability that the observation was generated from component h. To decide the class for the observation, the maximum a posteriori (MAP) estimate can be used. That is, the h corresponding to $\max_h(\kappa_h)$.

2.2 The special-case GMCM for meta-analysis

In the Li et al. [21] reproducibility analysis, the $p \times d$ matrix $[x_{gk}]$ consists of test-statistics or p-values interrogating the same null hypothesis for a large number p of e.g., genes for each of $d \geq 2$ studies. Rows corresponds to genes, indexed by g, and columns to experiments, indexed by k. Without loss of generality, *large* values are considered to be indicative of the alternative hypothesis. A prototypical example in genomics is a matrix of transformed p-values for the hypothesis of no differential expression of genes between treatment and control groups for two or more experiments. The task is here to determine which genes g are commonly significant in all experiments. Ordinary meta-analysis methodologies involve combining confidence intervals of effect sizes, test-statistics, or p-values in a row-wise manner and assessing the significance whilst controlling the number of false positives [25].

Li et al. [21] proposed a special case of Equation 4 with $m = 2$ components corresponding to whether the null or alternative hypothesis is true, where $h = 1$

corresponds to spurious signals and $h = 2$ to genuine ones. Hence α_1 and $\alpha_2 = 1 - \alpha_1$ is the fraction of spurious and genuine signals, respectively. Li et al. [21] further assumes the following constraints on the parameters

$$\begin{aligned} \boldsymbol{\mu}_1 &= \mathbf{0}_{d\times 1} \quad = (0, 0, \ldots, 0)^\top, \\ \boldsymbol{\mu}_2 &= \mathbf{1}_{d\times 1}\mu = (\mu, \mu, \ldots, \mu)^\top, \quad \mu > 0 \end{aligned} \tag{6}$$

and

$$\boldsymbol{\Sigma}_1 = \mathbf{I}_{d\times d} = \begin{bmatrix} 1 & 0 & \cdots \\ 0 & 1 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}, \qquad \boldsymbol{\Sigma}_2 = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \cdots \\ \rho\sigma^2 & \sigma^2 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}, \tag{7}$$

where $\rho \in [-(d-1)^{-1}, 1]$ and $\sigma^2 > 0$. The lower bound on ρ is a requirement for $\boldsymbol{\Sigma}_2$ to be positive semi-definite. In other words, if the null-hypothesis is true, the latent variable is a d-dimensional standard multivariate normal distribution. If not, it is an latent d-dimensional multivariate normal distribution with equal means and a compound symmetry covariance structure. Figure 1 (D-F) shows an example of the observed, copula, and latent levels of the special GMCM where $d = 2$.

With the above constraints the special model is parameterized by only $\boldsymbol{\theta} = (\alpha_1, \mu, \sigma^2, \rho)$ whereby the dimensionality of the parameter space is substantially reduced. Furthermore, all marginal cdfs are equal, $\Gamma_1 = \cdots = \Gamma_d$, and similarly are all pdfs equal, $\gamma_1 = \cdots = \gamma_d$.

Li et al. [21] defines the *local irreproducibility discovery rate* of an observation as

$$\operatorname{idr}(\mathbf{u}) = \kappa_1 = P(H = 1 \mid \mathbf{u}, \boldsymbol{\theta}), \tag{8}$$

analogously to the local false discovery rate (Lfdr) of Efron [10, 11, 12]. Notice, that Equation 5 coincide with Equation 8 for the special model. As the multiple testing problem is present when more observations are obtained, an adjusted *irreproducibility discovery rate* was also defined by Li et al. [21]:

$$\operatorname{IDR}(\alpha) = P(H = 1 \mid \mathbf{u} \in I_\alpha, \boldsymbol{\theta}), \tag{9}$$

where $I_\alpha = \{\mathbf{u} \mid \operatorname{idr}(\mathbf{u}) < \alpha\}$, i.e., the probability of a gene being non-reproducible while in the rejection region. The adjusted $\operatorname{IDR}(\alpha)$ relates to idr in the same manner as marginal false discovery rate (mFDR) relates to the Lfdr.

2.3 The GMCM likelihood function

Suppose we have observed p i.i.d. samples

$$\mathbf{x}_1 = (x_{11}, \ldots, x_{1d}), \ldots, \mathbf{x}_p = (x_{p1}, \ldots, u_{pd})$$

from Equation 4 which can be arranged into the observation matrix introduced in Section 2. From these, the marginal uniform variables $\mathbf{u}_1 = F_\circ(\mathbf{x}_1) = (u_{11}, \ldots, u_{1d}), \ldots, \mathbf{u}_p = F_\circ(\mathbf{x}_p) = (u_{p1}, \ldots, u_{pd})$ are computed and are independent and identically distributed according to the copula density of Equation 3. The log-likelihood is thus given by

$$\ell\big(\boldsymbol{\theta}; \{\mathbf{x}_g\}_{g=1}^p\big) \propto \ell\big(\boldsymbol{\theta}; \{\mathbf{u}_g\}_{g=1}^p\big) = \sum_{g=1}^p \log c(\mathbf{u}_g; \boldsymbol{\theta}) \tag{10}$$

$$= \sum_{g=1}^p \log \sum_{h=1}^m \frac{\alpha_h}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_h|}} \exp\left(-\frac{1}{2} \big(\Gamma_\circ^{-1}(\mathbf{u}_g) - \boldsymbol{\mu}_h\big)^\top \boldsymbol{\Sigma}_h^{-1} \big(\Gamma_\circ^{-1}(\mathbf{u}_g) - \boldsymbol{\mu}_h\big) \right)$$

$$- \sum_{g=1}^p \sum_{k=1}^d \log \sum_{h=1}^m \frac{\alpha_h}{\sqrt{2\pi\Sigma_{hkk}}} \exp\left(-\frac{1}{2\Sigma_{hkk}} \big(\Gamma_k^{-1}(u_{gk}) - \mu_{hk}\big)^2 \right),$$

since the Jacobian arising from transformation $F_\circ$ is not dependent on $\boldsymbol{\theta}$ (and thus constant when optimizing with respect to $\boldsymbol{\theta}$).

In practice, $F_1, \ldots, F_d$ are unknown and estimated by the empirical cdf

$$\hat{F}_k^{(p)}(x) = \frac{1}{p} \sum_{g=1}^p \mathbb{1}[x_{gk} \leq x].$$

Hence the pseudo-observations

$$\hat{u}_{gk} = \hat{F}_k^{(p)}(x_{gk}) = \frac{1}{p} \operatorname{rank}(x_{gk}) \tag{11}$$

of u_{gk} are plugged into the log-likelihood and the maximizing parameters are found. However, since p is large, $\hat{F}_k^{(p)}$ is a good estimate of F_k and thus $\hat{u}_{gk} = \hat{F}_k^{(p)}(x_{gk}) \approx F_k(x_{gk}) = u_{gk}$. The GMCM is rank-based since plugging a variable into its empirical cdf corresponds to a particular ranking scheme in which the lowest value is awarded rank 1 and ties are given their largest available rank. To avoid infinities in the computations $\hat{u}_{gk}$ is rescaled by the factor $\frac{p}{p+1}$.

The usage of $\hat{u}_{gk}$ violate the assumption of independent observations as the ranking introduces dependency between the observations. The introduced dependency is arguably negligible when p is large. We ignore this problem and refer to Chen et al. [6] and the references therein for a more detailed discussion about this problem which is common to all copula model estimation procedures.

3 The GMCM package

3.1 Package overview

The **GMCM** package currently have 14 user visible functions of which the majority are for convenience. The functions are presented in Table 1 and the

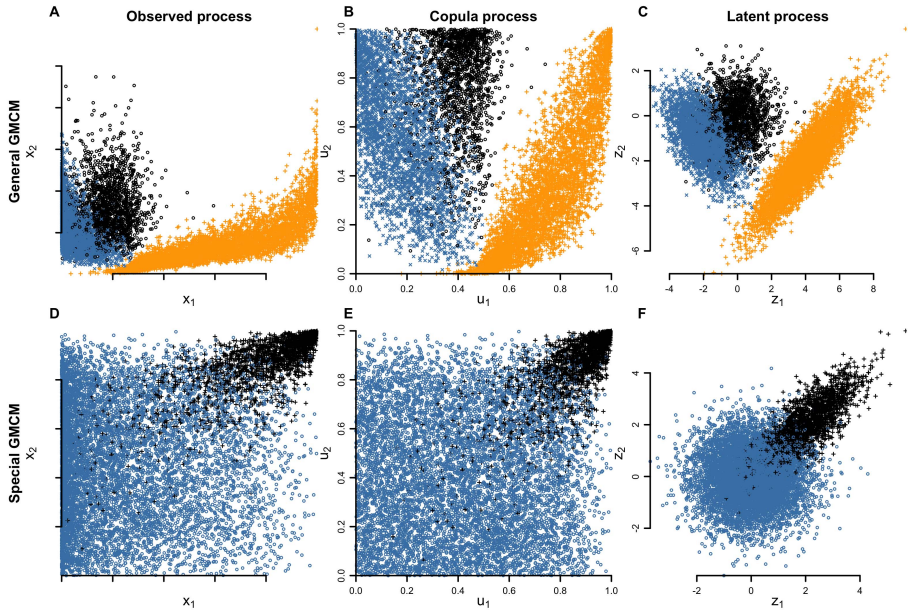


Fig. 1: From left to right the observed, copula (or rank), and latent process is shown. The first and second row of panels illustrate 10,000 realizations from the general and special model, respectively. The component from which the realizations come are visualized by colour and point-type. Each dimension in the special model corresponds to an experiment where simultaneously high values in both experiments are indicative of good reproducibility.

Table 1: Overview of the visible user functions and their purpose in approximate order of importance. Confer the documentation (e.g., `help("Uhat")`) for function arguments and return types. Relevant equations are right-justified.

Function	Description	
`fit.full.GMCM`	Fit the general model	(4)
`fit.meta.GMCM`	Fit the special model	(4)(6)(7)
`get.prob`	Get class probabilities for the general model	(5)
`get.IDR`	Get class probabilities (idr and IDR) for the special model	(8)(9)
`SimulateGMCMData`	Generate samples from a GMCM	(4)
`SimulateGMMData`	Generate samples from a GMM	(1)
`Uhat`	Rank and scale the columns of the argument.	(11)
`choose.theta`	Choose starting parameters in the general GMCM.	
`full2meta`	Convert from `theta`-format to `par`.	
`meta2full`	Convert from `par`-format to `theta`.	
`rtheta`	Generate random `theta`.	
`is.theta`	Test if `theta` is correctly formatted.	
`rmvnormal`	Generate multivariate gaussian observations.	
`dmvnormal`	Fast evaluation of multivariate Gaussian pdf.	

GMCM reference manual. Two different parameter formats are used depending on use of the special or general model. In the general model a specially formatted list of parameters is used, named `theta` in function arguments and documentation. The `rtheta` function generates such a prototypical list with random parameters and `is.theta` conveniently tests if the argument is properly formatted. If the special model is to be used, the required parameters are simply given in a numeric vector $(\alpha_1, \mu, \sigma, \rho)$ of length 4, named `par` in arguments and documentation. The useful functions `meta2full` and `full2meta` provide easy conversion between the general `theta` and the special `par` format.

The most important functions `fit.full.GMCM` and `fit.meta.GMCM` fit the general and special GMCMs, respectively. The `method` argument of these functions specify the optimization routine to be used. If the general model is used `get.prob` returns a matrix of posterior probabilities κ_{gk} defined in Equation 5. In the special model, the `get.IDR` is used to compute local idr (i.e., the posterior probability of belonging to the irreproducible component) and adjusted IDR values.

The `SimulateGMMData` and `SimulateGMCMData` functions provide simulation of observations from the models specified in Equations 1 and 4, respectively.

Beside the following tutorial, a small usage example of the special model is also found in `help("GMCM")`. All simulations and computations were carried out on a regular laptop (1.7 GHz Intel Core i5, 4GB DDR3 RAM).

3.2 Using the package

We proceed with a small tutorial on the package. As an illustration, we load the package and simulate 10,000 observations from a 2-dimensional 3-component GMCM with randomly chosen parameters in the following manner:

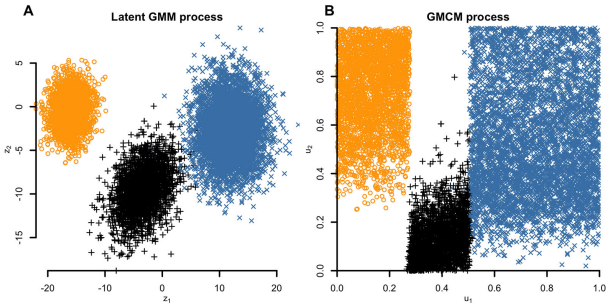


Fig. 2: Panel A shows realizations from the latent process and panel B the corresponding marginally uniformly distributed process. Note, that while B shows true realizations from the GMCM $\mathbf{u}_g$ the ranked observed values $\hat{\mathbf{u}}_g$ are almost visually identical because of the relative large number of observations.

```
> library("GMCM")
> set.seed(100)
> n <- 10000
> sim <- SimulateGMCMData(n = n, theta = rtheta(m = 3, d = 2))
```

The `sim` object is a `list` containing the `matrix` of the realized latent process (`sim$z`), the `matrix` of true realizations from the GMCM density (`sim$u`), the formatted parameters (`sim$theta`), and the component from which each observation is realized (`sim$K`). Figure 2 shows the realized data.

Subsequently, we select a starting estimate from the data, fit the ranked observed data using Nelder-Mead (`NM`), and compute the posterior probabilities of each observation belonging to each component:

```
> ranked.data <- Uhat(sim$u)
> start.theta <- choose.theta(ranked.data, m = 3)
> mle.theta <- fit.full.GMCM(u = ranked.data, theta = start.theta,
+                            method = "NM", max.ite = 10000,
+                            reltol = 1e-4)
> kappa <- get.prob(ranked.data, theta = mle.theta)
> Khat <- apply(kappa, 1, which.max)
```

The function `Uhat` ranks and rescales as described in Section 2.3. The `choose.theta` function uses the k-means algorithm on the rank-level to find an initial set of parameters. From the k-means clustering, crude estimates of the mixture proportions, mean values, and variances can be computed. The correlations in all components are taken to be zero. This usually provides reasonable initial parameters. Objections may be made to using such a procedure on the rank and not latent level. However, as we are only interested in the relative position of the components this often serves as a reasonable starting parameter. The `fit.full.GMCM` does the actual optimization of the likelihood

Table 2: Confusion matrices of GMCM and k-means clustering results.

	$\hat{H}$ **(GMCM)**			$\hat{H}$ **(k-means)**		
	1	2	3	1	2	3
H						
1	2747	0	0	2693	54	0
2	0	2276	6	5	2270	7
3	0	57	4914	26	882	4063

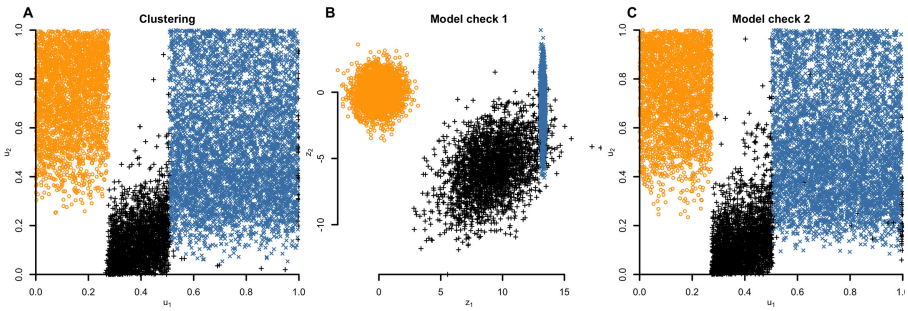


Fig. 3: Panel A shows the estimated class labels of the observations by colour and point-type. As a model control panel B and C shows 10,000 realizations from the GMM and GMCM using the fitted parameters.

to arrive at the MLE. The default Nelder-Mead (`NM`) procedure converged in 499 iterations in about 6 seconds.

In serious applications the starting values should be chosen carefully and the algorithm ought to be started at different positions of the parameter space to investigate the stability and uniqueness of the maximum likelihood estimate. The estimate with the largest likelihood should then be chosen.

The confusion matrix for the GMCM clustering, seen in Table 2, yields an accuracy of 99.4%. In (unfair) comparison, the k-means algorithm have an accuracy of 90.3%. Figure 3 shows the clustering results and simple model checks by simulation from the fitted parameters. Though a high clustering accuracy is achieved, we see from the model check in Figure 3B compared to Figure 2A that the underlying parameters are not really identifiable. However, we see from panel C, that the fitted parameters model the observed ranks closely and thus provide a high predictive accuracy.

3.3 Runtime and technical comparison

For the special model, the **GMCM** package implements an arbitrary number of dimensions (or experiments) d to be included whereas the **idr** package only supports $d = 2$. The **GMCM** package considerably decreases the per iteration runtime of the pseudo expectation-maximization (PEM) algorithm compared

Table 3: Runtime comparisons of the **idr** and **GMCM** packages with increasing number of observations p. The benchmarked optimization procedures are the pseudo EM algorithm (PEM) and the Nelder-Mead (NM) method. The runtime is given in seconds. The last column shows the relative speed per iteration compared to the fastest procedure.

p / Package	Algorithm	Runtime (s)	Iterations (n)	s/n	Rel. speed
1,000					
idr	PEM	3.03	22	0.138	50.4
GMCM	PEM	1.27	125	0.010	3.7
GMCM	NM	0.75	275	0.003	1.0
10,000					
idr	PEM	17.64	15	1.176	143.7
GMCM	PEM	4.16	163	0.025	3.1
GMCM	NM	1.94	237	0.008	1.0
100,000					
idr	PEM	257.63	17	15.155	304.2
GMCM	PEM	40.79	258	0.158	3.2
GMCM	NM	10.71	215	0.050	1.0

to the **idr** package. The optimization procedures such as Nelder-Mead (NM), simulated annealing (SANN), and others which only rely on evaluations of the likelihood further reduce the runtime compared to the PEM.

Run and iteration times for an increasing number of observations are seen in Table 3 on a simulated dataset with parameters $(\alpha_1, \mu, \sigma, \rho) = (0.7, 2, 1, 0.9)$. The algorithms were all run with the starting values $(0.5, 2.5, 0.5, 0.8)$. The parameters were chosen such that the **idr** package does not converge prematurely.

To assess the optimization routines in the **idr** and **GMCM** packages, 1000 datasets with 10,000 observations were simulated from the special model with parameters $\boldsymbol{\theta} = (0.9, 3, 2, 0.5)$. The special model was fitted to each of the datasets using each of the available routines with random initial parameter values. Figure 4 shows the results from the fitting procedures. The maximum number of iterations were set to 2,000. The SANN procedure was given 3,000 iterations.

The clusters of parameter estimates away from the true values seen in Figure 4 presumably corresponds to local maxima of the likelihood. Hence many of the procedures are fairly often caught in such local maxima. Interestingly, while the estimates of the standard deviation $\hat{\sigma}$ and correlation $\hat{\rho}$ for the PEM algorithm seem to be biased, the algorithm achieved a high clustering accuracy. We also see that the PEM algorithms in **GMCM** and **idr** behave quite differently. The maximal number of iterations, 2000, was hit only by the PEM algorithm 274 and 18 times for the **idr** and **GMCM** packages, respectively. Also notable is the factor 555 reduction in total runtime from between the fastest and slowest fitting procedures.

All warnings produced by the PEM algorithm in **idr** was `"NaNs produced"`. PEM in **GMCM** only warned that the maximum number of iterations was

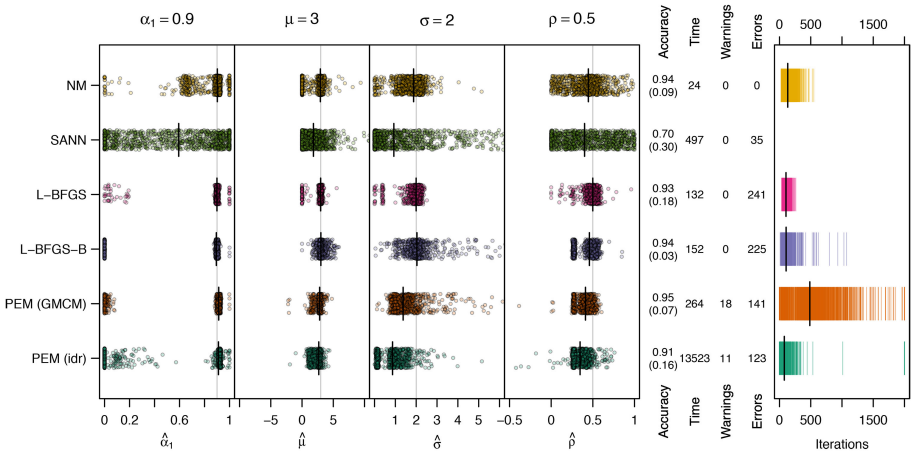


Fig. 4: Parameter fitting results for the different optimization procedures. From left to right, the first four panels show plots of the fitted parameter estimates. The true parameter values are plotted as vertical lines. Next, the mean clustering accuracy (and standard deviation), total run time in minutes for all 1000 fits, and the number of warnings and errors are shown. The last panel shows the number of iterations for each fit. The black vertical lines indicate the median.

reached. The errors produced by SANN and L-BFGS-B seemingly arise as the estimates of the covariance matrix became singular. The vast majority of the errors by L-BFGS were divergence to non-finite likelihood values. The only unique error thrown by PEM (**idr**), `"missing value where TRUE/FALSE needed"`, seems to stem from a simple bug.

Considering computational efficiency and robustness, accuracy, and precision of parameter estimates, we chose the Nelder-Mead as the default optimization procedure.

3.4 Availability of the package

The **GMCM** package is open-source and available both at the CRAN (Comprehensive R Archive Network) and at the GitHub repository `https://github.com/AEBilgrau/GMCM.git` for bug reports as well as easy forking and editing.

4 Maximum likelihood estimation

4.1 Maximizing the likelihood

The optimization of the likelihood function in Equation 10 is non-trivial. There exists no closed form expression for Γ_k^{-1}. Furthermore there are intrinsic problems of identifiability of the GMCM parameters. These problems will greatly affect any estimation procedure.

Both Li et al. [21] and Tewari et al. [33] make use of a pseudo EM (PEM) algorithm to find the maximizing parameters. Tewari et al. [33] use the PEM as a "burn-in" and switch to a gradient descent algorithm. Both authors derive the likelihood function of the GMM, ℓ_{GMM}, specified by Equation 1 and the estimators for the corresponding EM algorithm. The PEM algorithm then iteratively alternates between estimating pseudo-observations $\hat{z}_{gk} = \Gamma_k^{-1}(\hat{u}_{gk}; \boldsymbol{\theta})$ and subsequently updating $\boldsymbol{\theta}$ by an E and M step. While this intuitively is a viable approach, it effectively ignores the Jacobian of Equation 3 as the transformation Γ^{-1} depends on the parameters $\boldsymbol{\theta}$. In short, the wrong likelihood is thus optimized and a pseudo (or quasi) maximum likelihood estimate is found. This may yield an inefficient optimization routine and biased parameter estimates. This problem of the PEM is appreciated by Tewari et al. [33].

A fundamental problem with the PEM algorithm is the alternating use of pseudo-observations and parameter updates. The pseudo data is not constant in the ℓ_{GMM} which implies no guarantee of convergence nor convergence to the correct parameters.

To clarify, let $\boldsymbol{\theta}^{(m)}$ denote the m'th estimate of $\boldsymbol{\theta}$. From $\boldsymbol{\theta}^{(m)}$, pseudo data is estimated by

$$\hat{z}_{gk}^{(m)} = \Gamma_k^{-1}\big(\hat{u}_{gk}; \boldsymbol{\theta}^{(m)}\big), \quad g \in \{1, \dots, p\},\ k \in \{1, \dots, d\}.$$

The PEM algorithm alternates between updating parameter estimates and pseudo data which results in the following log-likelihood values,

$$\begin{aligned} \dots, \ell_{\text{GMM}}\big(\boldsymbol{\theta}^{(m)} \big| \{\hat{\mathbf{z}}_g^{(m)}\}_g\big),\ \ell_{\text{GMM}}\big(\boldsymbol{\theta}^{(m+1)} \big| \{\hat{\mathbf{z}}_g^{(m)}\}_g\big), \\ \ell_{\text{GMM}}\big(\boldsymbol{\theta}^{(m+1)} \big| \{\hat{\mathbf{z}}_g^{(m+1)}\}_g\big),\ \ell_{\text{GMM}}\big(\boldsymbol{\theta}^{(m+2)} \big| \{\hat{\mathbf{z}}_g^{(m+1)}\}_g\big),\ \dots, \end{aligned}$$

given in the order of computation. Conventionally, convergence is established when the difference of successive likelihoods is smaller than some $\epsilon > 0$. The implementation of Li et al. [21] through the package **idr** for R determines convergence if

$$\ell_{\text{GMM}}\big(\boldsymbol{\theta}^{(m+1)} \big| \{\hat{\mathbf{z}}_g^{(m+1)}\}_{g=1}^p\big) - \ell_{\text{GMM}}\big(\boldsymbol{\theta}^{(m)} \big| \{\hat{\mathbf{z}}_g^{(m)}\}_{g=1}^p\big) < \epsilon,$$

where $\epsilon > 0$ is pre-specified. However, an increase in successive likelihoods is only guaranteed by the EM algorithm when the (pseudo) data are constant. Since both the pseudo data and parameter estimate have changed the above difference can be, and often is to our experience, negative. In the **idr** package this sometimes happens in the first iteration without warning. Such cases arguably stops the procedure prematurely since a negative difference obviously is smaller than some positive ϵ. The EM algorithm only guarantees that the difference

$$\ell_{\text{GMM}}\big(\boldsymbol{\theta}^{(m+1)} \big| \{\hat{\mathbf{z}}_g^{(m)}\}_{g=1}^p\big) - \ell_{\text{GMM}}\big(\boldsymbol{\theta}^{(m)} \big| \{\hat{\mathbf{z}}_g^{(m)}\}_{g=1}^p\big)$$

is non-negative and thus might be more suitable for determining convergence.

The PEM convergence criterion used by Tewari et al. [33] is when the difference in successive parameters estimates is sufficiently small while recording the highest observed likelihood estimate which partly remedy the problem. However, the PEM still inherits the conventional problems of the EM algorithm. It often exhibit slow convergence and offers no guarantee for finding the global optimum.

Our software package **GMCM** offers fast optimization of both the general and special models. Our implementation of the PEM algorithm supports various convergence conditions. By default, it determines convergence when

$$\left|\ell_{\text{GMCM}}\left(\boldsymbol{\theta}^{(m+1)}\middle|\{\hat{\mathbf{u}}_g^{(m)}\}_{g=1}^p\right) - \ell_{\text{GMCM}}\left(\boldsymbol{\theta}^{(m)}\middle|\{\hat{\mathbf{u}}_g^{(m)}\}_{g=1}^p\right)\right| < \epsilon.$$

and returns the parameters which yield the largest likelihood. This is not necessarily the one obtained in the last iteration. The internal function `PseudoEMAlgorithm` is called when `fit.full.GMCM` or `fit.meta.GMCM` are run with `method = "PEM"`.

Instead of the EM approach, however, we propose optimizing the GMCM likelihood function in Equation 10 using procedures relying only on likelihood evaluations. To make this a feasible approach considerable effort has been put into evaluating the log-likelihood function of Equation 10 in a fast manner by implementing core functions in C++ using **Rcpp** and **RcppArmadillo** [8, 9, 14]. With fast likelihood evaluations the standard `optim` optimization procedure in R is used with various optimization procedures, such as Nelder-Mead (the amoeba method), simulated annealing, and BFGS quasi-Newton methods.

When the parameters are passed to `optim` we use various transformations to reformulate the optimization problem as an unconstrained one. We logit-transform the mixture proportions. In the general model, a Cholesky decomposition combined with a log-transformation is used to ensure positive definiteness of the covariance matrices. In the special model, the variance σ^2 is ensured positive by a log-transform. The restriction on the correlation ρ to the interval $[-(d-1)^{-1}, 1]$ is guaranteed by an affine and logit function composition.

Additional speed have also been obtained by fast inversion of the marginals Γ_k. Similarly to Li et al. [21], we linearly interpolate between function evaluations. However, we distribute the default 1,000 function evaluations to each component according to the current estimate of the mixture proportions. The determined number of function evaluations for component h within the k'th dimension is then sampled equidistantly in the interval $\mu_{hk} \pm a\sqrt{\Sigma_{hkk}}$ where $a = 5$ by default. Lastly, the monotonicity of Γ_k is used to quickly invert the function by reflection around the identity line. Furthermore, we approximate the mixture cdf Γ_k by using the approximation of the error function $\text{erf}(x) \approx 1 - (a_1 t + a_2 t^2 + a_3 t^3)\exp(-x^2)$ where $t = 1/(1 + bx)$ and a_1, a_2, a_3, and b are constants [1, p. 299; 16].

Table 4: Equivalent optima in pure noise. A dot ($\cdot$) denotes an arbitrary value. The given values need only to be approximate.

Situation	α_1	μ	σ	ρ
1	1	$\cdot$	$\cdot$	$\cdot$
2	0	$\cdot$	$\cdot$	0
3	$\cdot$	0	1	0

4.2 Identifiability of parameters

The model suffers from unidentifiable parameter configurations. As a consequence of the GMCM invariance to translations only relative distances between the location parameters $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_m$ can be inferred. We arbitrarily anchor the first component at $\boldsymbol{\mu}_1 = \mathbf{0}$ as a partial solution. To account for scaling invariance, the first component is required to have unit variance in each dimension, that is $\Sigma_{1kk} = 1$ for all k. However, problems of identifiability persists in a number of scenarios. In cases where two or more components in the latent GMM are well-separated from each other the relative distances and component variances are not identifiable for all practical purposes. For example in the special GMCM, the parameter configuration $\boldsymbol{\theta} = (0.5, 10, 1, 0)$, say, will be indistinguishable from $(0.5, 100, 1, 0)$. The ranking destroys all information about the relative variances and distances between the well-separated components.

The clustering might also easily fail when the location and variation parameters for two or more components are similar along the same dimension. Suppose for example that $\boldsymbol{\mu}_1 = (0, 0)$, $\boldsymbol{\mu}_2 = (4, 0)$, and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}_{2\times 2}$ where the location and variation parameters equal along the ordinate axis. In such cases, the ranking will create a homogeneous cluster which cannot be easily be separated.

Even though the parameters may not be fully estimable in all cases, the general model can still be an effective clustering algorithm if measured by clustering accuracy.

Table 4 describes three situations in the special model where the parameter estimates and thus the following clustering should be carefully interpreted. If the parameter estimate approaches any of the given numbers then the remaining parameters, represented by dots, are effectively non-identifiable. For example in Situation 1, if the mixture proportion α_1 approaches 1 then the remaining parameters can easily diverge as they no longer contribute to the likelihood. In Situation 2 where $\boldsymbol{\theta} = (0, \cdot, \cdot, \rho)$ extra caution should be displayed if ρ becomes substantially different from zero as all observations will be deemed reproducible. While the above corrections somewhat remedy these issues, the three situations can still be observed, especially when data consisting of nearly pure noise is supplied.

5 Applications

5.1 Reproducibility of microarray results

In molecular biology, microarrays are often used to screen large numbers of candidate markers for significant differences between case and control groups. Microarrays simultaneously probe the DNA composition or transcribed RNA activity of multiple genes in a biological sample. The number of probes ranges in the orders of 10,000 to 6,000,000, depending on the specific microarray.

In the study of haematological malignancies it is of biological interest to know how normal B-lymphocytes develop [18, 20, 30]. Hence, B-cells from removed tonsil tissue of six healthy donors were sorted and isolated using fluorescence-activated cell sorting (FACS) into five subtypes of B-cells: Naïve (N) B-cells, Centrocytes (CC), Centroblasts (CB), Memory (M) B-cells, and Plasmablasts (PB). As part of the immune response to an infection, the CBs proliferate rapidly and become CCs within the so-called germinal centres (GC). The 6×5 samples were profiled with *Affymetrix GeneChip HG-U133 plus 2.0* (U133) microarrays [See 3, for further details].

It is e.g., of interest to identify which gene expressions have been altered within the GCs from which the CCs and CBs come. We therefore tested the hypothesis of no difference in genetic expression between CC and CB samples against N, M, and PB samples for all the gene expressions present on the U133 array.

Since gene profiling technologies are rapidly evolving the experiment was later repeated with new donors and on the newer *GeneChip Human Exon 1.0 ST* (Exon) microarray.

The 30 samples on the U133 arrays and the 30 samples on Exon arrays were preprocessed and summarized to gene level separately and independently using the RMA algorithm with the R/Bioconductor package **affy** using custom CDF-files [7]. This preprocessing resulted in the genetic expression levels of 37,923 probe-sets for the U133 array and 19,750 probe-sets for the Exon array both annotated with Ensembl gene identifiers (ENSG identifiers).

Each experiment was analysed separately using a mixed linear model and empirical Bayes approach using the **limma** package [31] to test the hypothesis of no differential expression for each gene between the CC + CB and the N + M + PB groups. The tests yield two lists of p-values for the U133 and Exon arrays.

The p-value lists were reduced to the 19,577 common genes present on both array types and combined into a matrix $[x_{gk}]_{19577\times 2}$ where x_{gk} is one minus the p-value for varying gene expression for gene g in experiment $k \in \{\text{U133}, \text{Exon}\}$.

To determine the genes which are reproducibly differentially expressed, the special GMCM was fitted with the Nelder-Mead optimization procedure using `fit.meta.GMCM`. The procedure was started in 3 different starting values and the estimate with the largest log-likelihood was chosen. The best estimate converged in 311 iterations. Subsequently, the local and adjusted IDR values

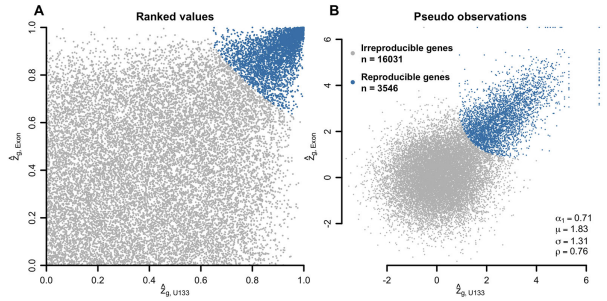


Fig. 5: Panel A shows a plot of the scaled ranks of p-values for the exon experiment against the scaled ranks of the p-values for the U133 experiment. Presumably, genes located in the upper left or lower right of the plots are false positive results in either experiment. Panel B shows the estimated latent GMM process. The fitted parameters shown are used to marginally transform panel A into the B.

were computed with `get.IDR`. A total of 3546 genes (18.1%) were found to have an adjusted IDR value below 0.05 and deemed reproducible. The results are illustrated in Figure 5 along with the parameter estimates. The algorithm successfully picks p-values which are high-ranking in both experiments.

If the MAP estimate, corresponding to a local idr value less than 0.5, is used then 4510 genes, corresponding to 23%, are deemed reproducible. This agrees with the estimate of the mixture proportion of the null component $\alpha_1 = 0.71$.

Note, since no biological ground truth is available, the accuracy cannot be determined. However, since genes which are not differentially expressed are expected to be irreproducible the accuracy may be high.

For comparison, the number of genes marginally significant at 5% significance level after Benjamini-Hochberg (BH) correction [2] is 3968 and 6713 for the U133 and Exon experiments, respectively. The number of commonly significant genes (i.e., simultaneously significant in both experiments) is 3140 or 16%. This corresponds to the common approach of using Venn diagrams.

The list of reproducible genes, which can be ranked by their idr-values, provides a more accessible list of genes for further biological down-stream analyses than the unordered list of genes obtained by the Venn diagram approach.

The p-values from the experiments are available in **GMCM** using `data("u133VsExon")`.

5.2 Effects of cryopreservation on reproducibility

Cryopreservation is a procedure for preserving and storing tissue samples by cooling them to sub-zero temperatures. It is convenient for researchers and a crucial component of biobanking. Cryopreservation is usually assumed by default to alter the biological sample since many cryopreserving substances are toxic, the freezing procedure may damaged the sample due to ice crystallization, and it may induce cellular stress response. Fresh is therefore consid-

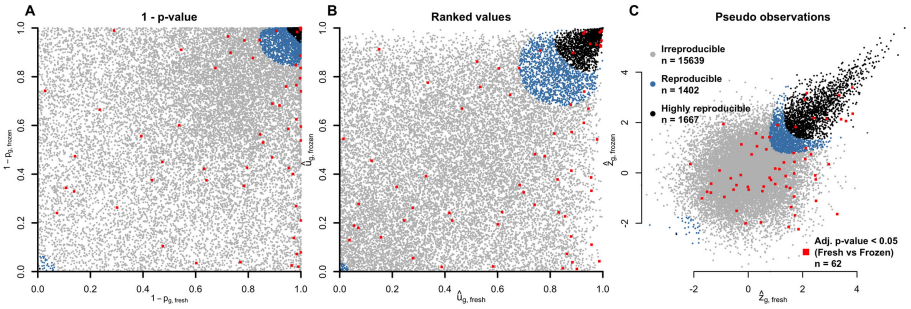


Fig. 6: Results from the reproducibility analysis of cryopreserved samples. Panel A shows the p-values p_g for the test of no differential expression between the pre- and post germinal centre groups for fresh and frozen samples. Panel B shows the corresponding ranked p-values $\hat{u}_g$, and panel C shows the estimated latent process $\hat{z}_g$. The estimated level of reproduciblity for each probe set is colour coded according to the legend in panel C. Genes significantly different across fresh and frozen samples are plotted as red squares regardless of the reproducibility level.

ered favorable to cryopreserved tissue. Few studies have analysed the effect of the cryopreservation on phenotyping and gene expression. Recently, we studied cryopreservation to gauge the actual impact of the cryopreservation on global gene expression in a controlled comparison of cryopreserved and fresh B-lymphocytes. Similarly to the above, the B-cells were prepared from peripheral blood of 3 individual healthy donors and FACS sorted into 2×4 B-cell subtypes, Immature (Im), Naïve (N), Memory (M), and Plasmasblasts (PB). Half of the samples were cryopreserved and thawed prior to the gene expression profiling using the Exon array while the other half was profiled fresh. The resulting data was preprocessed using RMA [See 26, for further details]. As a supplement to the manuscript, we performed a reproducibility analysis using the special model which however was omitted due to our concerns about complexity and added length to the manuscript.

If cryopreservation has relatively negligible effects on global screenings, then a high reproducibility should be expected for differential expression analyses within the fresh and frozen samples – however only for the true differentially expressed genes. For each probe set, the samples were analysed using linear mixed models as described in Rasmussen et al. [26] and the hypothesis of no differential expression between pre (Im + N) and post germinal centre (M + PB) cells was tested for both fresh and cryopreserved samples separately to mimic the situations where only fresh or frozen samples are available. The special GMCM was fitted using the resulting absolute value of the test-statistics to determine the level of reproducibility of each probe set. Local and adjusted irreproducible discovery rates were computed for all probe sets and this level of reproducibility was discretized into three groups: highly reproducible ($\text{IDR}_g < 0.05$, cf. Equation 9), reproducible ($\text{idr}_g < 0.5$, cf. Equation 8), and irreproducible ($\text{idr}_g \geq 0.5$).

The best parameter estimate of 40 fits was

$$\boldsymbol{\theta} = (\alpha_1, \mu, \sigma, \rho) = (0.73, 1.08, 1.32, 0.86)$$

. The reproducibility analysis deemed 1,667 (8.9%), 1,402 (7.5%), and 15,639 (83.6%) genes highly reproducible, reproducible, and irreproducible, respectively. Figure 6 shows these classifications of the p-values for differential expression between pre and post germinal cells for the fresh and frozen samples. The total of 3,069 (16.4%) reproducible probe sets seems quite high and agree with the estimated mixture proportion of 0.73. Again, the model correctly captures the genes with simultaneously low p-values. Recall also that non-differentially expressed genes are expected to be irreproducible and the actual accuracy is thus much higher although it (again) cannot easily be estimated when no biological ground truth is available.

Naturally, one might wonder whether genes changed due to cryopreservation to a large extent are deemed irreproducible. The paired design allowed us to investigate this hypothesis. The hypothesis of no difference in expression between fresh and frozen samples for each gene was therefore tested and the significant BH-adjusted p-values at the 5% level are highlighted in Figure 6. The expectation above was then tested using a test for non-zero Spearman correlation between the p-values and idr-values which yielded a non-significant correlation ($\rho = 0.009, p = 0.21$). In other words, high evidence for a change between fresh and frozen is not associated with greater irreproducibility (idr). Alternatively, a Fisher's exact test also did not yield a difference in odds (odds ratio $= 0.67$, 95 % CI $= (0.36, 1.32)$, p-value $= 0.23$) of having a BH-adjusted significant change due to cryopreservation in the reproducible group (odds $= 48/(15591 - 48)$) compared to the irreproducible (odds $= 14/(3055 - 14)$). Thus there is no evidence for an over-representation of the irreproducible genes among the significant one. We might thus conclude that though some genes change due to cryopreservation, the differential analysis between subgroups to a great extent still yields the same results whether the samples are fresh or frozen.

Lastly, notice that some genes in the lower-left of Figure 6 (A-C) near the origin are also being deemed reproducible. This is an artifact of the model due to the high correlation of $\rho = 0.86$ in the reproducible component.

The p-values and test scores are available in **GMCM** using `data("freshVsFrozen")`.

5.3 Image segmentation using the general GMCM

In computer vision and graphics, image segmentation is useful to simplify and extract features of pictures. To illustrate the flexibility of the model and the computational capability of the **GMCM** package a 1.4 Mpx (965×1500 px) image of the Space Shuttle Atlantis, seen at the top of Figure 7, was segmented into 10 colours.

The JPEG image can be represented as a $1{,}447{,}500 \times 3$ matrix where each column corresponds to a colour channel in the RGB colour space and each row corresponds to a pixel and observation in the GMCM. The values are in this case on the interval $[0, 1]$.

A 3-dimensional, 10-component GMCM was fitted using the PEM algorithm which resulted in the middle image of Figure 7. The segmented colours were chosen using the location estimates $\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_{10}$. That is, the three dimensional vector $\hat{F}_\circ^{-1}(\Gamma_\circ(\hat{\boldsymbol{\mu}}_h; \boldsymbol{\theta})) \in [0, 1]^3$ in the RGB space was used as the colour of cluster h. Alternatively, the average RGB value of each cluster could be used.

For comparison the 1.4 Mpx image was also segmented with the k-means algorithm. The results are seen at the bottom of Figure 7. The final colours given to each cluster was the means estimated by the algorithm.

As seen, the k-means and GMCM yield quite different segmentations and different details of the image are captured. For example, the GMCM seem to capture more details of the bottom of the orange external tank. However perhaps erroneously, the GMCM also cluster the black left edge of the photo together with a light cluster. The superior method is dependent on the application at hand. We acknowledge that disregarding spatial correlations between pixels is quite naïve. However, this example should illustrate the computational capability of the package of handling large datasets with a high number of clusters.

The package **jpeg** was used to read, manipulate, and write the JPEG image from R [35].

6 Concluding remarks

The software for the gradient descent algorithm used by Tewari et al. [33] to arrive at a maximum likelihood estimate is written in the proprietary language MATLAB but not provided as open source. Hybrid procedures, similar to the one proposed by Tewari et al. [33], can easily be constructed with the **GMCM** package. The **GMCM** package solves some of the previously described issues regarding the maximum likelihood estimation and provides a considerable speed-up in computation times. However, there seems to be no complete remedy for all of the challenges of the GMCMs. As stated, the transformation into uniform marginal distributions by ranking will result in a loss of information about the distance between components that are well separated.

The intrinsic identifiability problems of GMCMs may in practice often not be a big issue. Even though the parameters of the assumed underlying GMM can be difficult to estimate due to the flat likelihood function, the clustering accuracy can still be very high. Furthermore, the actual parameters, except perhaps the mixture proportions, does often not seem of particular interest in applications. Hence, the merit of the GMCMs should be measured by predictive accuracy which still remains to be explored. In this respect, we believe that the theoretical and practical properties of the special GMCM and IDR approach

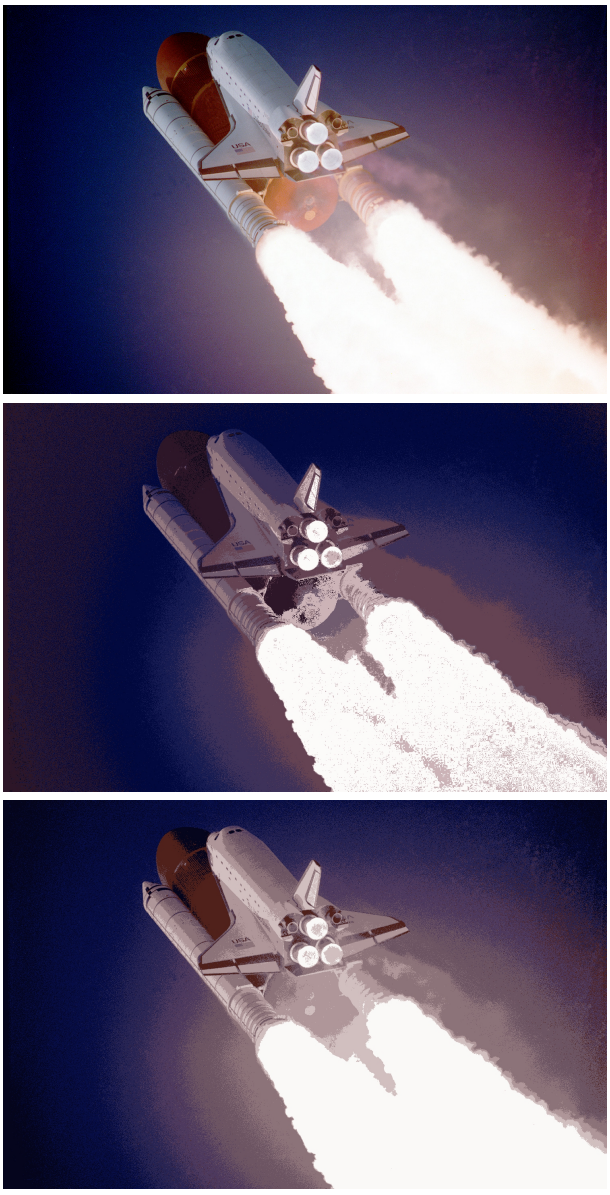


Fig. 7: Top: The original 1.4 Mpx JPEG image of the space shuttle Atlantis' climb to orbit during mission STS-27 in December 1988. Middle: The image segmented into 10 colours by the GMCM. Bottom: The image segmented into 10 colours by k-means clustering. Image credit: NASA.

should be studied further and compared to common p-value combining meta analyses, such as the methods of Fisher, Stouffer, Wilkinson, Pearson, and others, see e.g., Owen [25]. Interestingly and perhaps of slight concern, it can be seen that the IDR approach would be deemed unreasonable by Condition 1 in Birnbaum [5] whenever $\rho \neq 0$. It is unclear whether the method fulfills properties such as admissibility [5] and relative optimality in Bahadur's sense [22].

The simulation study in Section 3.3 revealed relatively many errors thrown by the **GMCM** package. We are committed to pinpoint the exact sources of the errors and provide fixes in future versions. We suspect the errors encountered are due to divergence of the parameters and should therefore be treated as such. With this in mind we believe that software should fail loudly with error or warning when it indeed fails.

In conclusion, the **GMCM** package provides a fast implementation of the flexible and widely applicable tool for reproducibility analysis and unsupervised clustering. The flexibility and applicability is however gained at the cost of a complicated likelihood function.

Acknowledgements

We thank Andreas Petri for his help on the microarray preprocessing workflow. The technical assistance from Alexander Schmitz, Julie S. Bødker, Ann-Maria Jensen, Louise H. Madsen, and Helle Høholt is also greatly appreciated. As are the helpful statistical comments from Steffen Falgreen. This research is supported by MSCNET, EU FP6, CHEPRE, the Danish Agency for Science, Technology, and Innovation as well as Karen Elise Jensen Fonden.

References

[1] M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions*. Dover Publishing Inc. New York, 1970. ISBN 0-486-61272-4.

[2] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(1):289–300, 1995.

[3] K. S. Bergkvist, M. Nyegaard, M. Bøgsted, A. Schmitz, J. S. Bødker, S. M. Rasmussen, M. Perez-Andres, S. Falgreen, A. E. Bilgrau, M. K. Kjeldsen, et al. Validation and implementation of a method for microarray gene expression profiling of minor B-cell subpopulations in man. *BMC Immunology*, 15(1):3, 2014.

[4] B. E. Bernstein, E. Birney, I. Dunham, E. D. Green, C. Gunter, and M. Snyder. An integrated encyclopedia of dna elements in the human

genome. *Nature*, 489(7414):57–74, Sept. 2012. ISSN 1476-4687. doi: 10.1038/nature11247.

[5] A. Birnbaum. Combining independent tests of significance. *Journal of the American Statistical Association*, 49(267):559–574, 1954.

[6] X. Chen, Y. Fan, and V. Tsyrennikov. Efficient estimation of semiparametric multivariate copula models. *Journal of the American Statistical Association*, 101(475):1228–1240, Sept. 2006. ISSN 0162-1459. doi: 10.1198/016214506000000311.

[7] M. Dai, P. Wang, A. D. Boyd, G. Kostov, B. Athey, E. G. Jones, W. E. Bunney, R. M. Myers, T. P. Speed, H. Akil, S. J. Watson, and F. Meng. Evolving gene/transcript definitions significantly alter the interpretation of genechip data. *Nucleic Acids Research*, 33(20):e175, Jan. 2005. ISSN 1362-4962. doi: 10.1093/nar/gni179.

[8] D. Eddelbuettel. *Seamless R and C++ Integration with* **Rcpp**. Springer-Verlag, New York, 2013. ISBN 978-1-4614-6867-7.

[9] D. Eddelbuettel and R. François. **Rcpp**: Seamless R and C++ Integration. *Journal of Statistical Software*, 40(8), 2011.

[10] B. Efron. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, 99(1): 96–104, 2004.

[11] B. Efron. Local false discovery rates. Technical report, Division of Biostatistics, Stanford University, 2005. URL http://statistics.stanford.edu/~ckirby/techreports/BIO/BIO234.pdf.

[12] B. Efron. Size, power and false discovery rates. *The Annals of Statistics*, 35(4):1351–1377, Aug. 2007. ISSN 0090-5364. doi: 10.1214/009053606000001460.

[13] L. Ein-Dor, O. Zuk, and E. Domany. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 103 (15):5923–8, Apr. 2006. ISSN 0027-8424. doi: 10.1073/pnas.0601231103.

[14] R. François, D. Eddelbuettel, and D. Bates. **RcppArmadillo***:* **Rcpp** *Integration for Armadillo Templated Linear Algebra Library*, 2012. URL http://CRAN.R-project.org/package=RcppArmadillo. R package version 0.3.6.1.

[15] F. E. Harrell, Jr. **Hmisc***: Harrell Miscellaneous*, 2014. URL http://CRAN.R-project.org/package=Hmisc. R package version 3.14-3.

[16] C. Hastings, J. T. Hayward, and J. P. Wong. *Approximations for Digital Computers*, volume 170. Princeton University Press Princeton, 1955.

[17] J. P. Ioannidis, E. E. Ntzani, T. a. Trikalinos, and D. G. Contopoulos-Ioannidis. Replication validity of genetic association studies. *Nature genetics*, 29(3):306–9, Nov. 2001. ISSN 1061-4036. doi: 10.1038/ng749.

[18] R. Küppers. Mechanisms of B-cell lymphoma pathogenesis. *Nature Reviews Cancer*, 5(4):251–262, Apr. 2005. ISSN 1474-175X. doi: 10.1038/nrc1589.

[19] F. Leisch. `Sweave`: Dynamic generation of statistical reports using literate data analysis. In W. Härdle and B. Rönz, editors, *Compstat 2002 — Proceedings in Computational Statistics*, pages 575–580. Physica Verlag, Heidelberg, 2002. ISBN 3-7908-1517-9.

[20] G. Lenz and L. M. Staudt. Aggressive lymphomas. *The New England Journal of Medicine*, 362(15):1417–1429, 2010. ISSN 1533-4406. doi: 10.1056/NEJMra0807082.

[21] Q. Li, J. B. J. Brown, H. Huang, and P. J. Bickel. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics*, 5 (3):1752–1779, 2011. ISSN 1932-6157. doi: 10.1214/11-AOAS466.

[22] R. Littell and J. Folks. Asymptotic optimality of fisher's method of combining independent tests. *Journal of the American Statistical Association*, 66(336):802–806, 1971.

[23] R. Nelsen. *An Introduction to Copulas*. Springer-Verlag, 2nd edition, 2006. ISBN 978-0387286594.

[24] E. Neuwirth. **RColorBrewer***: ColorBrewer Palettes*, 2011. URL `http://CRAN.R-project.org/package=RColorBrewer`. R package version 1.0-5.

[25] A. B. Owen. Karl pearson's meta-analysis revisited. *The Annals of Statistics*, 37(6B):3867–3892, Dec. 2009. ISSN 0090-5364. doi: 10.1214/09-AOS697.

[26] S. M. Rasmussen, A. E. Bilgrau, A. Schmitz, S. Falgreen, K. S. Bergkvist, A. M. Tramm, J. Bæch, C. L. Jacobsen, M. Gaihede, M. K. Kjeldsen, J. S. Bødker, K. Dybkær, M. Bøgsted, and H. E. Johnsen. Stable phenotype of B-cell subsets following cryopreservation and thawing of normal human lymphocytes stored in a tissue biobank. *Cytometry Part B: Clinical Cytometry*, 2014. ISSN 1552-4957. doi: 10.1002/cytob.21192.

[27] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

[28] Revolution Analytics. **doMC***: Foreach Parallel Adaptor for the* **multicore** *Package*, 2014. URL `http://CRAN.R-project.org/package=doMC`. R package version 1.3.3.

[29] Revolution Analytics and S. Weston. **foreach***: Foreach Looping Construct for R*, 2014. URL http://CRAN.R-project.org/package=foreach. R package version 1.4.2.

[30] L. Rui, R. Schmitz, M. Ceribelli, and L. M. Staudt. Malignant pirates of the immune system. *Nature Immunology*, 12(10):933–40, Oct. 2011. ISSN 1529-2916. doi: 10.1038/ni.2094.

[31] G. K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1):1–25, 2004.

[32] P. Tan, T. Downey, and E. S. Jr. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Research*, 31 (19):5676–5684, Oct. 2003. ISSN 1362-4962. doi: 10.1093/nar/gkg763.

[33] A. Tewari, M. J. Giering, and A. Raghunathan. Parametric characterization of multimodal distributions with non-gaussian modes. *ICDM 2011 conference*, pages 286–292, Dec. 2011. doi: 10.1109/ICDMW.2011.135.

[34] The Encode Consortium. A User's Guide to the Encyclopedia of DNA Elements (ENCODE). *PLoS Biology*, 9(4):e1001046, Apr. 2011. ISSN 1545-7885. doi: 10.1371/journal.pbio.1001046.

[35] S. Urbanek. **jpeg***: Read and Write JPEG Images*, 2012. URL http://CRAN.R-project.org/package=jpeg. R package version 0.1-2.

[36] Y. Xie. *Dynamic Documents with* R *and* **knitr**. CRC Press, 2013. ISBN 9781482203530.

[37] M. Zhang, C. Yao, Z. Guo, J. Zou, L. Zhang, H. Xiao, D. Wang, D. Yang, X. Gong, J. Zhu, Y. Li, and X. Li. Apparently Low Reproducibility of True Differential Expression Discoveries in Microarray Studies. *Bioinformatics*, 24(18):2057–63, Sept. 2008. ISSN 1367-4811. doi: 10.1093/bioinformatics/btn365.

Paper II

Stable Phenotype of B-cell Subsets Following Cryopreservation and Thawing of Normal Human Lymphocytes Stored in a Tissue Biobank

Anders Ellern Bilgrau*, Simon Mylius Rasmussen*, Alexander Schmitz, Steffen Falgreen, Kim Steve Bergkvist, Anette Mai Tramm, John Bæch, Chris Ladefoged Jacobsen, Michael Gaihede, Malene Krag Kjeldsen, Julie Støve Bødker, Karen Dybkær, Martin Bøgsted, and Hans Erik Johnsen
*Shared first authorship

Published in
Cytometry Part B: Clinical Cytometry Vol. 88(1), pp. 40–49, 2015.

Preface: This applied paper substantiates a basic premise of cryopreservation. Cryopreservation conserves biological samples by cooling them to sub-zero temperatures and halting their chemical reactivity. As such, it is an indispensable component in bio-banking and convenient as it enables analyzing biological samples in batches. The premise is that cryopreserved samples are largely unchanged compared to fresh samples. Hence, this application investigates which gene expressions (and cell populations sizes) differs between fresh and frozen B-cell subsets. Alternatively, it can be framed as a study of reproducibility: *Do analyses carried out on frozen samples yield the same results as those on fresh samples?*

The work is important for Dybkær et al. [11] which uses both fresh and frozen samples.

One obvious critique of the paper comes back to the old statistical adage that one *fails to reject* and not accept the null hypothesis—or, absence of evidence is not evidence of absence. In this paper, the lack of significance differences between the fresh and frozen samples might well simply be a lack of statistical power. This was one reason we performed the extra reproduciblity analysis included in Section 5.2 of Paper I.

 The layout has been revised.

Stable Phenotype of B-cell Subsets Following Cryopreservation and Thawing of Normal Human Lymphocytes Stored in a Tissue Biobank

Abstract

Background Cryopreservation is an acknowledged procedure to store vital cells for future biomarker analyses. Few studies, however, have analyzed the impact of the cryopreservation on phenotyping.

Methods We have performed a controlled comparison of cryopreserved and fresh cellular aliquots prepared from individual healthy donors. We studied circulating B-cell subset membrane markers and global gene expression, respectively by multiparametric flow cytometry and microarray data. Extensive statistical analysis of the generated data tested the concept that "overall, there are phenotypic differences between cryopreserved and fresh B-cell subsets". Subsequently, we performed a consecutive uncontrolled comparison of tonsil tissue samples.

Results By multiparametric flow analysis, we documented no significant changes following cryopreservation of subset frequencies or membrane intensity for the differentiation markers CD19, CD20, CD22, CD27, CD38, CD45, and CD200. By gene expression profiling following cryopreservation, across all samples, only 16 out of 18708 genes were significantly up or down regulated, including FOSB, KLF4, RBP7, ANXA1 or CLC, DEFA3, respectively. Implementation of cryopreserved tissue in our research program allowed us to present a performance analysis, by comparing cryopreserved and fresh tonsil tissue. As expected, phenotypic differences were identified, but to an extent that did not affect the performance of the cryopreserved tissue to generate specific B-cell subset associated gene signatures and assign subset phenotypes to independent tissue samples.

Conclusions We have confirmed our working concept and illustrated the usefulness of vital cryopreserved cell suspensions for phenotypic studies of the normal B-cell hierarchy; however, storage procedures need to be delineated by tissue specific comparative analysis.

1 Introduction

Glycerol as a cryoprotective agent for gametes was introduced in 1949 [23]. Since then the storage of cells in liquid nitrogen has evolved into a frequently used and well-documented technique for human tissue. Modern standard cryopreservation procedures for hematopoietic cells include a rate controlled freezing in dimethyl sulfoxide (DMSO) and storage of vital cells in the vapor phase of liquid nitrogen until thawing and use in therapy [1, 7]. Recent improvements in technology have increased our understanding of cellular and molecular events involved in cancer pathogenesis. Translation of such knowledge into clinical work requires the establishment of biobanks with access to stored biomaterial

and standardized technologies before validation and implementation of new biomarkers into patient care. Cryopreservation of collected single cell tissue samples from patients suffering from hematological disorders is convenient for subsequent scientific validation experiments. However, it is crucial that the cryopreserved tissue after thawing reflects the corresponding fresh samples in order not to introduce analytical bias.

Our research interest lies in the identification and characterization of the normal and malignant B-cell hierarchy including studies of the pre- and post germinal center transcriptional gene regulations and the cell of origin for diagnostic classification [17, 19]. However, we have recognized that studies of the potential biological impact of cryopreservation are insufficient and do not take into consideration the impact of the still evolving analytic technologies. There is a need to move from the simple storage of samples to a new level of sample handling and analysis complexity including high quality processed samples with associated analytic data and phenotypic characteristics to enable development of tools to make use of them. This point to the future of a well organized biostatistical activity to provide access and tools for data analysis to be stored in the biobank associated data base. This is illustrated by the present controlled comparative study including an extensive statistical evaluation of the analytic data with the purpose to investigate the impact of long or short term cryopreservation. The data available were from multiparametric flow cytometry (MFC) based identification of surface expressed membrane markers and the microarray (MA) based global gene expression profiling (GEP) of sorted B-cell subsets from cryopreserved and fresh peripheral blood, prepared in parallel from individual healthy donors. Our expectation was that the outcome of the statistical analysis most likely would identify significant phenotypic differences between cryopreserved and fresh B-cell subsets, however, of minor impact that would allow for the generation of standard operating procedures for storage of high quality biomaterial.

2 Material and Methods

2.1 Donor blood and tonsil tissue

All blood samples were collected following the national guidelines for transfusion medicine, by the local blood bank, Aalborg University Hospital and tonsil tissue samples were collected in accordance with the research protocol MSC-NET (Myeloma Stem Cell Network), accepted by the local ethical committee (N-20080062MCH). Biological material included into the study was peripheral blood from a total of 9 healthy donors and tonsil tissue obtained from routine tonsillectomies of 17 healthy donors as previously described [19]. Mononuclear cells (MNC) were isolated following Ficoll-Paque Plus (GE Health Care, Uppsala, Sweden) gradient centrifugation in accordance with the manufacturer's instructions and used immediately or following cryopreservation as described.

2.2 Multiparametric flow cytometry (MFC) analysis

Peripheral blood from 6 healthy donors was analyzed by FCM in parallel of fresh and cryopreserved cells. For each of the 6 blood samples approximately 3×106 MNC were suspended in $100\mu l$ phosphate-buffered saline (PBS) including 2% fetal bovine serum and stained by a direct immunofluorescence technique. Samples were incubated for 30 minutes on ice in the dark in an analytic 5-color combination set up with the following monoclonal antibody panel: $4\mu l$ per test of CD20 clone 2H7 conjugated with pacific blue (ExBio, Vestec, Czech Republic), $5\mu l$ per test of CD45 clone HI30 conjugated with Pacific Orange (Invitrogen, Denmark), $5\mu l$ per test of CD38 clone HB7 conjugated with fluorescein isothiocyanate (BD Biosciences, San Jose, CA), $20\mu l$ per test of CD22 clone S-HCL-1 conjugated with phycoerythrin (BD Biosciences, San Jose, CA), $5\mu l$ per test of CD27 clone O323 conjugated with peridinin chlorophyll protein/cyanine 5.5 (BioLegend, San Diego, CA), $5\mu l$ per test of CD19 clone SJ25C1 conjugated with phycoerythrin/cyanine 7 (BD Biosciences, San Jose, CA), $5\mu l$ per test of CD200 clone OX104 conjugated with allophycocyanin (eBioscience, San Diego, CA).

Tonsils from 17 healthy donors were analyzed as either fresh or cryopreserved samples of 1-3 $\times$ 106 MNCs incubated for 30 minutes on ice, in the dark, and in an 8-color combination set up [19] with the following monoclonal antibody panel: $20\mu l$ per test of CD20 clone 2H7 conjugated with pacific blue (eBioscience, San Diego, CA), $5\mu l$ per test of CD45 clone 2D1 conjugated with Anemonia majano cyan (AmCyan; BD Biosciences, San Jose, CA), $10\mu l$ per test of CD3 clone SK7 conjugated with fluorescein isothiocyanate (BD Biosciences), $20\mu l$ per test of CXCR4 clone 12G5 conjugated with phycoerythrin (Beckman Coulter, Brea, CA), $5\mu l$ per test of CD44 clone IM7 conjugated with peridinin chlorophyll protein/cyanine 5.5 (eBioscience), $10\mu l$ per test of CD10 clone HI10a conjugated with phycoerythrin/ cyanine 7 (BD Biosciences), $10\mu l$ per test of CD27 clone L128 conjugated with allophycocyanin (BD Biosciences), and $10\mu l$ per test of CD38 clone HIT2 conjugated with Alexa Fluor 700 (AF700) (ExBio, Vestec, Czech Republic). Following incubation the cells were centrifuged ($500g$, 5 min) and the cell pellet was washed and resuspended once with 3 ml of staining buffer by centrifugation ($250g$, 10 min). Finally, cells were resuspended in 0.5 ml of stain buffer for analysis. Before analysis the cytometer went through a setup and tracking application in the BD FACSDiva software according to the manufacturer's instructions. Samples were acquired by using FACSDiva software (BD Biosciences) on a 3-laser (488, 633, and $405nm$) FACS Canto II (BD Biosciences, San Jose, CA) [18]. Compensation was automatically calculated by the FACSDiva software using single stained control samples with the monoclonal antibodies previously listed.

2.3 Fluorescence-activated cell sorting (FACS)

Peripheral blood from 3 healthy donors was analyzed as paired samples of fresh and cryopreserved samples. MNC were sorted following the preparation procedure as for immunophenotypic characterization but stained with the following panel per 3 × 106 MNC/100μl staining volume: 10μl per test of CD20 clone 2H7 conjugated with pacific blue (eBioscience, San Diego, CA), 5μl per test of CD45 clone 2D1 conjugated with Anemonia majano cyan (AmCyan; BD Biosciences, San Jose, CA), 5μl per test of CD10 clone HI10a conjugated with phycoerythrin cyanine 7 (BD Biosciences, San Jose, CA), 20μl per test of CD19 clone SJ25C1 conjugated with peridinin chlorophyll protein/cyanine 5.5 (BD Biosciences, San Jose, CA), 3μl per test of CD27 clone L128 conjugated with allophycocyanin (BD Biosciences, San Jose, CA) and 4μl per test of CD38 clone HIT2 conjugated with Alexa Fluor 700 (AF700) (ExBio, Vestec, Czech Republic). Samples were acquired and sorted by using FACSDiva software (BD Biosciences) on a 3 laser (488, 633, and 405nm) FACSAria (BD Biosciences, San Jose, CA) [18]. Before the actual sorting, a total of 50,000 events per tube were acquired to define the gating regions and sort a minimum of 10,000 B-cells per subset as defined below.

2.4 Identification and enumeration of B-cell subsets

Immediately after staining, 1 × 106 events per test tube were acquired for i) scatter properties to eliminate debris, doublets, and dying cells and ii) fluorescent signals to define MNC leukocytes and B-cell subsets as delineated below.

- For the peripheral blood FCM analysis set ($N = 6$):
  - Immature (Im): CD45+, CD19+, CD20+, CD27−, CD38++, CD200+, CD22dim/+;
  - Naive(N): CD45+, CD19+, CD20+, CD27−, CD38−, CD200+, CD22+;
  - Memory (M): CD45+, CD19+, CD20+, CD27+, CD38−, CD200−/dim, CD22+;
  - Plasmablasts (PB): CD45+, CD19+, CD20−, CD27++, CD38++, CD200−, CD22−.
- For the peripheral blood paired FACS set ($N = 3$):
  - Immature (Im): CD45+, CD19+, CD20+, CD10+, CD27−, CD38++;
  - Naive (N): CD45+, CD19+, CD20+, CD10−, CD27−, CD38−;
  - Memory (M): CD45+, CD19+, CD20+, CD10−, CD27+, CD38−;
  - Plasmablasts (PB): CD45+, CD19+, CD20−, CD10−, CD27++, CD 38++.
- For the tonsils FCM analysis and FACS set ($N = 17$):
  - Naive (N): CD20+, CD44+, CD10−, CD27−, CD38−;
  - Memory (M): CD20+, CD44+, CD10−, CD27+, CD38−;
  - Plasmablasts (PB): CD20+, CD44+, CD10−, CD27++, CD 38++;
  - Centroblasts (CB): CD20+, CD44dim+, CD10+, CD38+, CD27het, CXCR4+;
  - Centrocytes (CC): CD20+, CD44dim+, CD10+, CD38+, CD27het, CXCR4−.

For enumeration, the leukocyte fraction was identified based on CD45 positivity and cell scatter properties. Owing to the higher fluorescence intensity, the CD19+ for peripheral blood and CD20+ expression for the tonsils were used to delineate the overall number of total B-cells for a minimum of 106 leukocytes per tube used for quantitation of B-cell subsets by a minimum of 100 positive events i.e. a counting accuracy of coefficient of variation $< 10\%$.

2.5 Freezing and thawing procedure

For the 9 analyzed peripheral blood samples, each individual had 6 vials with 107 MNC, 2 were analyzed fresh, 4 were cryopreserved and 2 were thawed after 24 hours as well as 2 after 336 hours (2 weeks). For the 9 tonsils analyzed, each sample with a minimum of 5 vials with 107 MNC were cryopreserved and thawed at varying storage time (weeks to months). From each sample, the vials of 107 MNC in $1ml$ suspension of 10% fetal bovine serum were cryopreserved by adding $0.5ml$ of 20% dimethyl sulfoxide (DMSO). Inversion of the vial ensured homogenous mixing of cells and cryopreservation medium and the samples were then subjected to decrease in temperature from room temperature to $-196°C$ by a controlled rate liquid nitrogen freezer (Planer Biomed, Sunbury-on-Thames, UK) and finally stored in liquid nitrogen. Thawing of the cryopreserved cells was done in a $37°C$ water bath until no ice clumps were detectable and then added to a $37°C$ mixture of RPMI-1640 medium with 30% fetal bovine serum and 1% antibiotics. The cells were then resuspended and centrifuged ($400g$, 5 min) at room temperature and the pellet was added to the medium before staining, analysis and sorting as described above.

2.6 Microarray (MA) procedures for analysis of blood and tonsil subsets

Circulating B-cell subsets were sorted in lysis/binding buffer (Miltenyi Biotech, Bergisch-Gladbach, Germany) and stored at $-20°C$ as previously described [6]. Messenger RNA (mRNA) was isolated using the μMACS technology (Miltenyi Biotech, Bergisch-Gladbach, Germany) and eluate was concentrated by a volume reduction step using a speedVac Concentrator 5310 (Eppendorf, Hamburg, Germany). For exon MA analysis, mRNA was converted to cDNA using polyT and random priming and amplified with Ovation Pico WTA system (NuGEN Technologies, Inc., San Carlos, CA). After the amplification the cDNA product was purified using the QIAquick PCR purification kit (QIAGEN, Hilden, Germany) and the yield and purity was measured on the NanoDrop (Thermo Fisher, Wilmington, CA). In addition, the quality of the amplified cDNA product was analyzed with the RNA 6000 Nano LabChip (Agilent technologies, Inc., Palo Alto, CA). Three μg of amplified and purified cDNA (antisense-sense orientation) was converted to sense transcript cDNA (ST-cDNA) using the WT-OvationTM Exon module (NuGEN Inc.). Five μg of ST-cDNA was fragmented and biotin labelled using EncoreTM Biotin Module (NuGEN Inc.) and hy-

bridized to GeneChip Human Exon 1.0 ST Arrays (Affymetrix). However, one frozen PB sample was not hybridized due to insufficient ST-cDNA. Following an $18h$ hybridization step at $45°C$, the array was washed and stained according to the standard procedure (fluidic protocol FS450 0001, Affymetrix.com).

Tonsil B-cell subsets [6] were sorted directly into lysis buffer and mRNA was isolated and amplified using μMACS mRNA isolation kit (Miltenyi Biotechnologies, Bergisch Gladbach, DE) and WT-Ovation Pico RNA Amplification System (NuGEN, San Carlos, CA, USA), respectively, according to manufacturer's description. Cryopreserved and fresh samples were prepared for hybridization to GeneChip HG-U133 Plus 2.0 (Affymetrix) and Human Exon 1.0 ST Arrays (Affymetrix), respectively.

Processed arrays for all samples were scanned at $532nm$ using the GeneChip Scanner 3000 7G and CEL-files were generated by Affymetrix GeneChip Command Console Software (AGCC).

2.7 Analysis of MFC and MA data

For the MFC data analysis we used the Infinicyt software (Cytognos SL, Salamanca, Spain) [28]. Statistical analysis for gene expression was performed using Bioconductor packages [14] which are add-on modules for the statistical software R [24].

The effect of cryopreservation on frequencies of B-cell subsets by MFC was graphically assessed using correlation plots and Bland-Altman limits of agreement analysis [8]. A linear mixed model (LMM) with logit-transformed fractional numbers as outcome, cell subpopulation and type of cryopreservation as fixed effects, and donor as random effect, was used to test the null hypothesis of no differences in the logit transformed fractional numbers comparing fresh and $24h$ as well as $336h$ cryopreservation. These hypotheses were tested for both individual and pooled cell subsets. The pooled effect is a mean of Im, N, M, PB, and B subsets and is used to study the main effect of cryopreservation while the test within the individual subsets investigates interaction effects between subsets and cryopreservation. The model was fitted by the **lme4**-package [2, 21]. The contrasts of interest were computed by the **multcomp**-package [16]. The assumptions of the model of the population frequencies were checked by appropriate residual plots [21].

The MA data from blood were background corrected, quantile normalized, and summarized into 18,708 genes using the R/Bioconductor package **aroma.affymetrix** [3, 4] following the recent guidelines of Rodrigo-Domingo et al. [25]. As above, for MA data correlation plots and Bland-Altman limits of agreement analysis were made to evaluate the agreement between the different sample preparations for comparison. LMMs were fitted, analogously to the above, for each gene to test the null hypothesis of no differential $\log_2$ expression between fresh and cryopreserved samples for the 'pooled' and individual subsets with cell subset and type of cryopreservation as fixed effects and a random donor effect. To control the false-positive rate the P-values were

adjusted by the Benjamini-Hochberg procedure [5].

The B-cell subset associated gene signatures (BAGS) were generated from the cryopreserved tonsil samples as previously described [17] and subsequently used to classify the fresh samples. To do this we had to convert the GeneChip HG-U133 Plus 2.0 array to the probe sets of the HuEx 1.0 ST v2 array based on the file U133PlusVsHuEx_Complex.txt downloaded from Affymetrix.com. For each probe set used in the classifier the conversion was performed as an average of the probe sets on the fresh tonsil HuEx 1.0 ST v2 array data with a percent match above 90 using the percent match as weights. The GEP data were then probe set-wise centered to have a median of zero and scaled to have variance equal to the cryopreservation data on the HG-U133 plus 2.0 array. The result of the classification was assessed by the classification accuracy.

3 Results

In order to investigate the usefulness of vital cryopreserved cell suspensions for phenotypic studies of the normal B-cell hierarchy as an alternative to fresh processed samples, we performed a controlled comparison of cellular aliquots prepared from individual healthy persons by comparing frequency and membrane intensity as well as global gene expression for CD marker defined B-cell subsets in peripheral blood. Thereby, we attempt to minimize unwanted variation in the procedure by a) using the same donors, b) the same analysis and sorting equipment, c) the same antibodies and reagents and sorting strategy, and d) parallel handling of the paired samples after sorting for RNA-extraction and subsequent steps. Subsequently, we compared cryopreserved/thawed and fresh non-paired tonsil tissue to validate the use of stored biobank samples in biomarker research.

3.1 Immunophenotyping of circulating B-cell subsets

The total B-cell compartment defined as CD45+/CD19+ MNC were enumerated in fresh samples with frequencies having a median of 6% (range 4–12%) of which the Im subset had a median of 7% (range 7–9%), the N subset a median of 69% (range 56–78%), the M subset a median of 16% (range 6–25%), and the PB subset a median of 1% (range 0.2–2%). A statistical comparison of the B-cell subset frequencies for fresh, $24h$, and $336h$ stored MNC samples was performed with the results given in Table 1. The analysis by LMM documented no difference for Im, N, M, PB, or total B-cells. These results were in agreement with a visual inspection of cryopreserved/thawed and fresh subset sizes by the logit-transformed subset sizes as illustrated in Figure 1 and a high correlation (> 0.9) between cryopreserved and fresh subset frequencies, which were observed in Supplementary Figure 5A–B. Finally, the level of agreement is seen in Supplementary Figure 5C–D by corresponding Bland-Altman plots. Weak evidence (unadjusted P-value of 0.07 in Table 1) was seen for the seem-

Table 1: Comparison of fresh, 24, and 336 h stored cryopreserved (Froz24 and Froz336, respectively) samples of peripheral blood MNC used for enumeration by MFC immunophenotyping ($N = 6$). The linear mixed model (LMM) documented no significant differences at the 5% significance level. The pooled category is a combination of the Im, N, M, PB, and B-cells which corresponds to a main effect of cryopreservation. The individual subsets correspond to interactions between cryopreservation and that subset.

	Mean diff.	95% CI	z	P	Adj. P
Pooled					
Froz24 - Fresh = 0	−0.030	(-0.33,0.27)	−0.20	0.840	1.00
Froz336 - Fresh = 0	−0.031	(-0.33,0.27)	−0.21	0.840	1.00
Immature					
Froz24 - Fresh = 0	0.250	(-0.41,0.92)	0.75	0.460	1.00
Froz336 - Fresh = 0	0.190	(-0.47,0.85)	0.56	0.580	1.00
Naive					
Froz24 - Fresh = 0	−0.043	(-0.71,0.62)	−0.13	0.900	1.00
Froz336 - Fresh = 0	0.072	(-0.59,0.74)	0.21	0.830	1.00
Memory					
Froz24 - Fresh = 0	0.052	(-0.61,0.72)	0.15	0.880	1.00
Froz336 - Fresh = 0	−0.090	(-0.75,0.57)	−0.27	0.790	1.00
Plasmablast					
Froz24 - Fresh = 0	−0.600	(-1.3,0.063)	−1.80	0.076	0.91
Froz336 - Fresh = 0	−0.200	(-0.86,0.47)	−0.58	0.560	1.00
B-cells					
Froz24 - Fresh = 0	0.190	(-0.48,0.85)	0.55	0.580	1.00
Froz336 - Fresh = 0	−0.130	(-0.79,0.53)	−0.38	0.700	1.00

ingly systematic drop for PBs from fresh to 24 h frozen in Figure 1, which is accentuated in Supp. Figure 5C. The membrane marker fluorescence intensity (FI) was studied in a step by step comparison of the FI measured for each CD specific marker with only minor differences identified as illustrated in Figure 2A–D. Overall, FI's of CD45, CD19, CD20, CD27, CD38, CD22, and CD200 differ slightly, but not substantially, comparing corresponding cryopreserved vs. fresh samples.

3.2 Global gene expression of circulating B-cell subsets

The comparison of gene expression data for cryopreserved/thawed versus fresh samples ($N = 3$) across sorted subsets are given in Table 2, presenting significantly up- and down-regulated genes with a fold-change above 2 (i.e. $\log_2(\text{FC}) > 1$) for the LMM. A group of a priori selected B-cell specific genes of differentiation markers (CD) and transcription factors (TF) were inspected by the LMM z-scores for differential expression across the 4 subsets and gave the results shown in Supplementary Figure 6. The analysis identified 4 genes affected in the Im subsets, 14 genes in N, 13 genes in M, and 3 genes for the PB subsets. These

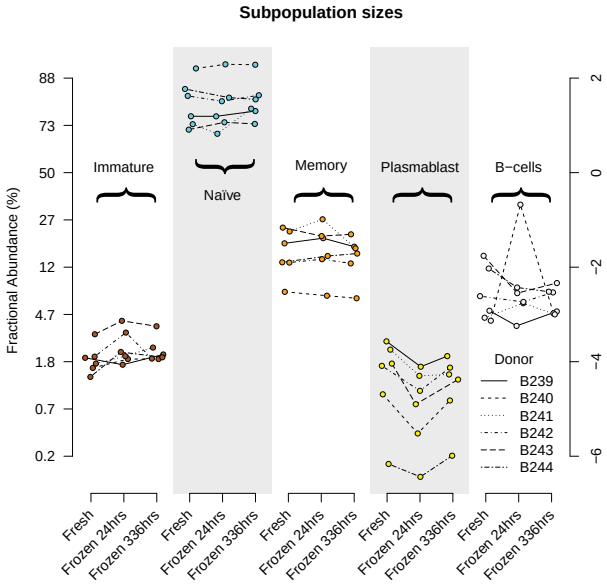


Fig. 1: Comparison of estimated frequencies for circulating B-cell subsets. Plot of the logit-transformed fractional sizes of the cryopreserved, and fresh blood B-cell subsets. Immature, Naive, Memory, Plasmablasts, and total B-cells reveal no substantial differences. The donor identity (ID B239–B244) for each of the six blood samples is also shown. The axes on the left and right, respectively, show the percentages on the original and logit scale.

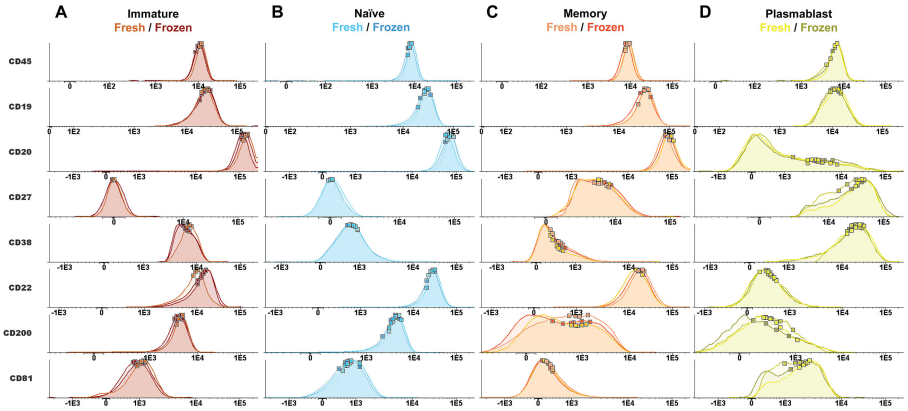


Fig. 2: A–D: Comparison of membrane CD intensity of circulating B-cell subsets. Graphical comparison of the fluorescent intensities of normal circulating B-cell population defining CD markers in fresh versus $24h/336h$ cryopreserved subsets from the peripheral blood FCM analysis set ($N = 6$), based on CD45, CD19, CD20, CD27, and CD38 expression supplemented with CD22, CD200, and CD81. Corresponding populations (fresh vs. frozen) are displayed as software overlaid and normalized densities with the mean values for each single of the 6 samples indicated as a square. The color codes for each fresh, $24h$, and $336h$ cryopreserved subsets are: Fresh (light color tone), $24h$ frozen (medium color tone), $336h$ frozen (dark color tone).

Table 2: Following cryopreservation we identified 16 out of the total number of 18,708 genes to be significantly different across all sample. This table gives the 11 and 5 significantly up- or down-regulated genes, respectively, from the LMM model analysis with a fold change above 2. Presented is the probe set ID (Probe set ID), the mean $\log_2$ expression value for the cryopreserved (μ_{fro}) and fresh (μ_{fre}) samples, the estimated log fold change (logFC) and its estimated standard error (SE), the 95% confidence interval (CI), and the Benjamini-Hochberg corrected P-value (Adj. P). DEF* refers to DEFA3/DEFA1/DEFA1B.

	Frozen vs. Fresh						
	ID	μ_{fro}	μ_{fre}	logFC	SE	CI	Adj. P
Up-regulated							
FOSB	3836266	8.66	7.08	1.51	.159	(1.19,1.82)	$5.71{\cdot}10^{-17}$
FOS	3544525	9.65	8.05	1.49	.197	(1.11,1.88)	$2.07{\cdot}10^{-10}$
KLF4	3219215	8.09	6.72	1.32	.204	(0.92,1.72)	$3.24{\cdot}10^{-07}$
LGALS2	3960174	9.45	7.75	1.64	.318	(1.01,2.26)	$4.62{\cdot}10^{-04}$
RBP7	2319550	7.82	6.64	1.07	.213	(0.65,1.49)	$6.99{\cdot}10^{-04}$
ANXA1	3174816	9.64	8.16	1.44	.302	(0.85,2.03)	$1.68{\cdot}10^{-03}$
CPVL	3043648	7.75	6.59	1.09	.245	(0.61,1.57)	$6.03{\cdot}10^{-03}$
ID2	2468622	7.44	6.07	1.27	.314	(0.65,1.88)	$2.43{\cdot}10^{-02}$
CSTA	2638869	9.79	8.46	1.26	.314	(0.65,1.88)	$2.43{\cdot}10^{-02}$
CLEC7A	3444009	7.47	6.19	1.23	.308	(0.63,1.83)	$2.76{\cdot}10^{-02}$
MS4A6A/MS4A4E	3374934	8.57	7.23	1.28	.324	(0.65,1.92)	$2.84{\cdot}10^{-02}$
Down-regulated							
CLC	3862108	4.72	6.94	−2.22	.375	(-2.96,-1.49)	$8.71{\cdot}10^{-06}$
DEF*	3122763	4.02	7.49	−3.43	.697	(-4.80,-2.06)	$9.74{\cdot}10^{-04}$
DEF*/DEFA11P	3122805	4.17	7.58	−3.36	.687	(-4.71,-2.01)	$9.97{\cdot}10^{-04}$
DEF*	3122784	4.41	7.62	−3.17	.653	(-4.45,-1.90)	$1.08{\cdot}10^{-03}$
TAS2R42	4053709	4.35	5.32	−1.02	.257	(-1.53,-0.52)	$2.76{\cdot}10^{-02}$

tests were, however, not adjusted for multiple testing. Most importantly, the genes KLF4 and c-jun were recurrently up-regulated by the freezing/thawing procedure compared to the fresh samples. Note, KLF4 also appear in Table 2. In general, a high correlation (> 0.9) between cryopreserved/thawed and fresh RMA normalized gene expressions were observed, as illustrated in Supplementary Figure 7A–D. The level of agreement was documented by the corresponding Bland-Altman analysis as illustrated in Supplementary Figure 8A–D. The percentage of genes outside the 99% limits of agreement was 2.6, 2.6, 2.1, and 2.5% for Im, N, M, and PB, respectively. The figure also shows that the defensin (DEF) gene family appear frequently with high fold-changes. Again, this may not be surprising as the DEF genes also appear in Table 2. In summary, the statistical analyses support our concept, but only a minor number of genes have changed expression following cryopreservation.

3.3 Implementation and performance of cryopreserved tonsil tissue

In our studies of normal B-cell subset in tonsil tissue, we have generated analytic data comparing cryopreserved biobank material ($N = 9$) and fresh tissue ($N = 8$) from consecutive but unpaired samples. The total B-cell compartment was defined as CD45+/CD20+/CD3- MNC and enumerated in cryopreserved

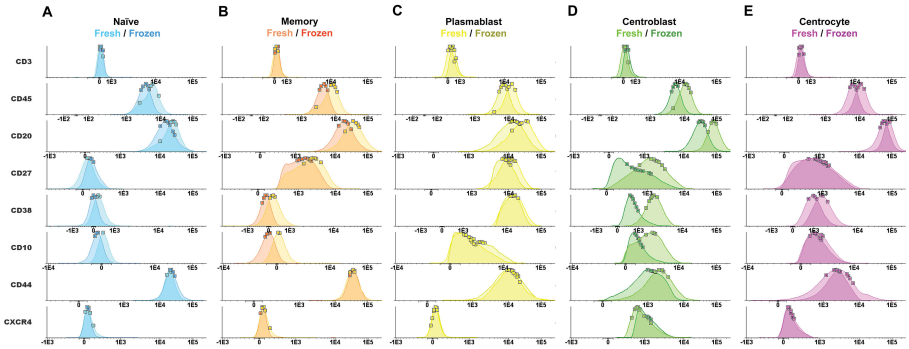


Fig. 3: A–E: Comparison of membrane CD intensity of tonsil tissue B-cell subsets. Graphical comparison of the fluorescent intensities of B-cell subpoplation defining CD markers in fresh versus cryopreserved normal tonsil tissue B-cell subsets, based on CD45, CD20, CD27, and CD38 expression as backbone markers supplemented with markers for tissue specific B-cells like CD44 and CXCR4, and the exclusion marker CD3. Corresponding merged populations (fresh vs. frozen) are displayed as software normalized and overlaid histograms for each CD marker with the population mean values for each single sample indicated as a square. The color codes for each set of fresh/cryopreserved subsets are: Fresh (light color tone), fro- zen (dark color tone).

and fresh samples with frequencies of median 74% (range 63–77%) and 57% (range 39–64%), respectively. However, the enumeration of the defined B-cell subsets in cryopreserved versus fresh samples did not differ: first, estimated for N to be of median 59% (range 46–63%) and 55% (range 43–63%); second, estimated for germinal center cells to be of median 18% (range 14–27%) and 17% (range 4–26%) including CB being either of median 1.5% (range 1–4%) and 5% (range 2-13%) or for CC a median of 15% (range 11–26%) and 12% (range 1–15%), respectively. Finally, we estimated for M a median of 21% (range 17–29%) and 26% (range 18–31%) and for PB median of 1% (range 1–4%) and 0.5% (range 0.2–1.2%), respectively. In a step by step comparison of the membrane FI measured for each CD specific marker only minor differences were identified as illustrated in Figure 3A–E. Overall, FI's of CD45, CD20, CD27, CD38, and CD10 differ slightly comparing corresponding cryopreserved vs. fresh samples, most prominent, however, for the CB subset (Figure 3A). Comparison of the defined subsets by multiparametric analysis of data gave the results as illustrated in Supplementary Figure 9A–B. In brief, the single tube data files were merged by the 'Infinicyt software' as described elsewhere [28]. Supplementary Figure 9A illustrates the best PCA separation view of the B-cell clusters composing the five B-cell subsets and their relation to each other based on the contribution of each fluorescence parameter to the separation of the clusters; and in comparison to the classical CD38 versus CD27 illustration for the classical B-cell subsets as defined by CD38 and CD27 expression from N, M, PB, and CB as well as CC illustrated in Supplementary Figure 9B.

To evaluate the performance of cryopreserved cell suspensions in MA studies of global gene expression, we analyzed the transcription level of the selected CD

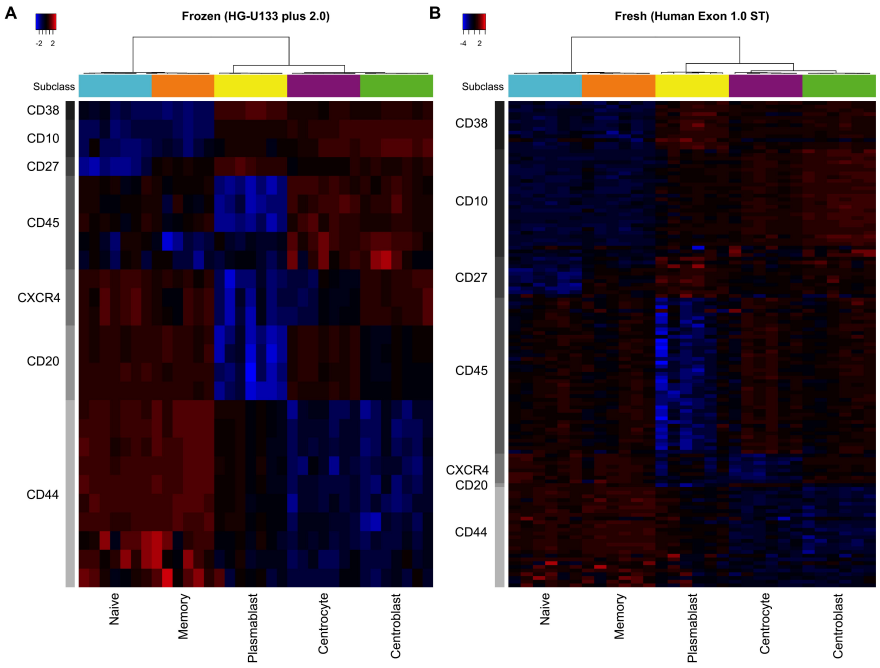


Fig. 4: A, B: The transcriptional level of the selected CD markers used to identify and sort the tonsillar B-cell subsets. Heat maps of the transcription level of CD marker probe sets and hierarchical cluster analysis for FACS sorted cryopreserved (A) samples on Affymetrix GeneChip HG-U133 Plus 2.0 and fresh (B) samples on the Human Exon 1.0 ST arrays. The mapped clusters illustrate that there is no substantial changes in the transcription level of the CD membrane marker due to cryopreservation or use of array platform. Each row represents one probe set and each column a separate B-cell subset as defined and isolated by cell sorting (A: $N = 8$ and B: $N = 6$). The heat map is color coded according to the $\log_2$ gene expression level (red is high, blue is low) as shown in the color key.

markers used to identify and sort each of the B-cell subsets. The hierarchical cluster analysis is illustrated in Figure 4A-B for sorted cryopreserved (A) and fresh (B) samples. The clustering validates the performance as expected with a substantial concordant robust transcription level expression of the membrane markers. In parallel the BAGS classifier generated by cryopreserved sample analysis was applied on the fresh samples and showed a perfect classification accuracy of 1.00 (with classification probabilities being in the range of 0.88–1.00). In summary, comparison of cryopreserved and fresh cell suspensions from tonsils revealed membrane and gene expression with differences of no major impact for tissue phenotyping, which will allow implementation of cryopreserved stored samples in studies of the normal B-cell hierarchy.

4 Discussion

By collecting and storing well-characterized sets of patient tissue, the same samples can be used in many different projects, reduce inconvenience to the donors, and allow more effective use of research biomaterial and funds. Studies of standardized storage of different tissues in biobanks can be continually extended as technologies and assays improve or more information becomes available about variables of particular interest in specific diseases. In this way comprehensive information can accumulate for well characterized and standard collected samples, which will ultimately increase the value of any subsequent future investigation. One fundamental problem with biobanking is to determine which samples should be collected and how they should be stored. Studies of somatic genetic changes that may have arisen during life, for example in tumor tissue, naturally require nucleated cells from the relevant tissue to be sampled. Consequently, collection and cryopreservation of single cell samples from normal donors and patients suffering from hematological disorders are a demand for future scientific experiments. However, in order not to introduce biological or technical bias it is crucial that the cryopreserved tissue after thawing reflects the corresponding fresh samples and most important potential tissue specific storage induced alterations are studied and delineated.

It is evident that the reinfusion of cryopreserved bone marrow cells reconstitutes the hematopoiesis indicating that no major damage of engrafting stem or early progenitor cells are introduced by cryopreservation. However, cryopreservation has been shown to cause chemical and physical cell stress due to the exposure to cryoprotectant as DMSO, and/or to the rate-controlled freezing and/or thawing procedure itself. A few publications have suggested that such sub-lethal damage may result in detectable changes in the expression of certain membrane markers associated with acute leukemia measured by flow cytometry underestimating some variables like CD5 and CD23 positivity in chronic lymphatic leukemia [9, 10]. The observed changes seem not only to be due to a selective loss of cells, but rather to unknown alterations of fragile surface expressed molecules that mechanically break off, are biological internalization upon stress or stable anchored into the membrane [12, 13, 15, 22, 26]. These findings are in accordance with our findings and it should be kept in mind that not all markers or B-cell subsets are equally robust when retrospective immunophenotypic studies of cryopreserved tissue are planned.

In parallel, it has been indicated by in vitro stimulation studies of activated cells that the functional potential of cryopreserved compared to fresh cells are equal by assessing the kinetics of proliferation, cell viability, cytokine, and membrane marker expression [20, 27]. In the present study, our primary goals were to describe the impact of cryopreservation on membrane markers and genes in normal blood B-cell subsets. In a controlled strategy, we compared fresh and cryopreserved aliquots from individual donor cells analyzed by MFC and GEP attempting to identify significant changes to challenge our concept that there are consequences of cryopreservation on B-cell subset membrane marker and

gene expression. Following an extensive statistical approach, our main findings were that circulating B-cells had only minor if any common traits that vary between cryopreserved and fresh cells, concluding that storage of nucleated cells induce a risk for inaccurate estimation of minor subset compartments and membrane molecule intensities—depending on some specific CD marker. Also, we identified single transcripts in nucleated blood cells including FOSB, KLF4, RBP7, ANXA1, CLC and DEFA3 at risk for cryospecific deregulation, which therefore should be avoided in biomarker studies in blood. These minor common traits have allowed us to implement the use of cryopreserved samples in our studies of normal B-cells and the performances illustrated for tonsil derived B-cell subsets characterized by the combination of MFC, FACS and GEP in our attempt to generate novel B-cell associated gene signatures [17, 19].

So far, our interpretation of the present results leads us to conclude that controlled vital cryopreserved cell samples can be used for future studies of subset specific gene signatures of the normal B-cell hierarchy.

Our conclusion may help in defining guidelines and recommendations for optimal tissue collection and storage but also for optimal interpretation of the gene expression results. To guarantee the highest possible quality of banked tissue samples and analytic databases in the future, each component of the activities and procedures needs to be studied in a quality assessment strategy generating standard operating procedures for tissue preparation, storage, analysis, and data handling—to obtain high quality biomaterial of clinical value.

5 Acknowledgements and Funding

The scientific program was supported by grants from the Danish Research Agency (grant no. 99 00 771, no. 271-05-0286, no. 271-05-0537, no. 22-00-0314 and no. 2101-07-0007), the Multiple Myeloma Research Foundation (senior grant 2003-4, contract no. 14), the EU 6th FP (grant no LSHC-CT-2006-037602 to MSCNET, coordinator HE Johnsen), 'The Obelske Family Foundation', 'The Heinrich Kopps Legat', and 'The Spar Nord Foundation'.

The authors have no conflict of interest to declare.

References

[1] A. M. Bakken. Cryopreserving human peripheral blood progenitor cells. *Current Stem Cell Research & Therapy*, 1(1):47–54, 2006.

[2] D. Bates, M. Maechler, B. Bolker, and S. Walker. **lme4***: Linear Mixed-Effects Models using Eigen and S4*, 2013. URL `http://cran.r-project.org/package=lme4`.

[3] H. Bengtsson, R. Irizarry, B. Carvalho, and T. P. Speed. Estimation and

assessment of raw copy numbers at the single locus level. *Bioinformatics*, 24(6):759–767, 2008.

[4] H. Bengtsson, P. Wirapati, and T. P. Speed. A single-array preprocessing method for estimating full-resolution raw copy numbers from all affymetrix genotyping arrays including GenomeWideSNP 5 & 6. *Bioinformatics*, 25 (17):2149–2156, 2009.

[5] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(1):289–300, 1995.

[6] K. S. Bergkvist, M. Nyegaard, M. Bøgsted, A. Schmitz, J. S. Bødker, S. M. Rasmussen, M. Perez-Andres, S. Falgreen, A. E. Bilgrau, M. K. Kjeldsen, et al. Validation and implementation of a method for microarray gene expression profiling of minor B-cell subpopulations in man. *BMC Immunology*, 15(1):3, 2014.

[7] D. Berz, E. M. McCormack, E. S. Winer, G. A. Colvin, and P. J. Quesenberry. Cryopreservation of hematopoietic stem cells. *American journal of hematology*, 82(6):463–472, 2007.

[8] J. M. Bland and D. G. Altman. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, 8(2):135–160, 1999.

[9] E. Costa, C. E. Pedreira, S. Barrena, Q. Lecrevisse, J. Flores, S. Quijano, J. Almeida, M. del Carmen Garcia-Macias, S. Bottcher, J. van Dongen, et al. Automated pattern-guided principal component analysis vs expert-based immunophenotypic classification of B-cell chronic lymphoproliferative disorders: A step forward in the standardization of clinical immunophenotyping. *Leukemia*, 24(11):1927–1933, 2010.

[10] V. Deneys, V. Thiry, N. Hougardy, A. Mazzon, P. Leveugle, and M. De Bruyère. Impact of cryopreservation on b cell chronic lymphocytic leukaemia phenotype. *Journal of immunological methods*, 228(1):13–21, 1999.

[11] K. Dybkær, M. Bøgsted, S. Falgreen, J. S. Bødker, M. K. Kjeldsen, A. Schmitz, A. E. Bilgrau, Z. Y. Xu-Monette, L. Li, K. S. Bergkvist, M. B. Laursen, M. Rodrigo-Domingo, S. C. Marques, S. B. Rasmussen, M. Nyegaard, M. Gaihede, M. B. Møller, R. J. Samworth, R. D. Shah, P. Johansen, T. C. El-Galaly, K. H. Young, and H. E. Johnsen. A diffuse large B-cell lymphoma classification system that associates normal B-cell subset phenotypes with prognosis. *Journal Of Clinical Oncology*, 33(12): 1379–1388, 2015.

[12] E. W. Fiebig, D. K. Johnson, D. F. Hirschkorn, C. C. Knape, H. K. Webster, J. Lowder, and M. P. Busch. Lymphocyte subset analysis on frozen whole blood. *Cytometry*, 29(4):340–350, 1997.

[13] K. R. Fowke, J. Behnke, C. Hanson, K. Shea, and L. M. Cosentino. Apoptosis: A method for evaluating the cryopreservation of whole blood and peripheral blood mononuclear cells. *Journal of Immunological Methods*, 244(1):139–144, 2000.

[14] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, 2004.

[15] A. Hansen, K. Reiter, T. Dörner, and A. Pruss. Cryopreserved human b cells as an alternative source for single cell mRNA analysis. *Cell and Tissue Banking*, 6(4):299–308, 2005.

[16] T. Hothorn, F. Bretz, and P. Westfall. Simultaneous inference in general parametric models. *Biometrical Journal*, 19(3):346–363, 2008. URL http://www.ncbi.nlm.nih.gov/pubmed/18481363.

[17] H. E. Johnsen, K. S. Bergkvist, A. Schmitz, M. K. Kjeldsen, S. M. Hansen, M. Gaihede, M. A. Nørgaard, J. Bæch, M.-L. Grønholdt, F. S. Jensen, et al. Cell of origin associated classification of B-cell malignancies by gene signatures of the normal B-cell hierarchy. *Leukemia & lymphoma*, 55(6): 1251–1260, 2014.

[18] T. Kalina, J. Flores-Montero, V. Van Der Velden, M. Martin-Ayuso, S. Böttcher, M. Ritgen, J. Almeida, L. Lhermitte, V. Asnafi, A. Mendonca, et al. Euroflow standardization of flow cytometer instrument settings and immunophenotyping protocols. *Leukemia*, 26(9):1986–2010, 2012.

[19] M. Kjeldsen, M. Perez-Andres, A. Schmitz, P. Johansen, M. Bøgsted, M. Nyegaard, M. Gaihede, A. Bukh, H. Johnsen, A. Orfao, and K. Dybkær. Multiparametric flow cytometry for identification and fluorescence activated cell sorting of five distinct B-cell subpopulations in normal tonsil tissue. *American Journal of Clinical Pathology*, 136(6):960–969, 2013.

[20] R. Mallone, S. Mannering, B. Brooks-Worrell, I. Durinovic-Belló, C. Cilio, F. S. Wong, and N. Schloot. Isolation and preservation of peripheral blood mononuclear cells for analysis of islet antigen-reactive T-cell responses: Position statement of the T-cell workshop committee of the immunology of diabetes society. *Clinical & Experimental Immunology*, 163(1):33–49, 2011.

[21] J. Pinheiro and D. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer Science & Business Media, 2006. URL http://CRAN.R-project.org/package=nlme.

[22] L. A. Pinto, M. T. Trivett, D. Wallace, J. Higgins, M. Baseler, M. Terabe, I. M. Belyakov, J. A. Berzofsky, and A. Hildesheim. Fixation and cryopreservation of whole blood and isolated mononuclear cells: Influence of different procedures on lymphocyte subset analysis by flow cytometry. *Cytometry Part B: Clinical Cytometry*, 63(1):47–55, 2005.

[23] C. Polge, A. Smith, A. Parkes, et al. Revival of spermatozoa after vitrification and dehydration at low temperatures. *Nature*, 164(4172):666, 1949.

[24] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL http://www.R-project.org/. ISBN 3-900051-07-0.

[25] M. Rodrigo-Domingo, R. Waagepetersen, J. S. Bødker, S. Falgreen, M. K. Kjeldsen, H. E. Johnsen, K. Dybkær, and M. Bøgsted. Reproducible probe-level analysis of the Affymetrix Exon 1.0 ST array with R/Bioconductor. *Briefings in Bioinformatics*, 15(4):519–533, 2014.

[26] M. Rosillo, F. Ortuno, J. Rivera, J. Moraleda, and V. Vicente. Cryopreservation modifies flow-cytometric analysis of hemopoietic cells. *Vox Sanguinis*, 68(4):210–214, 1995.

[27] S. Sattui, C. de la Flor, C. Sanchez, D. Lewis, G. Lopez, E. Rizo-Patrón, A. C. White, and M. Montes. Cryopreservation modulates the detection of regulatory T-cell markers. *Cytometry Part B: Clinical Cytometry*, 82 (1):54–58, 2012.

[28] J. van Dongen, L. Lhermitte, S. Böttcher, J. Almeida, V. Van der Velden, J. Flores-Montero, A. Rawstron, V. Asnafi, Q. Lecrevisse, P. Lucio, et al. Euroflow antibody panels for standardized n-dimensional flow cytometric immunophenotyping of normal, reactive and malignant leukocytes. *Leukemia*, 26(9):1908–1975, 2012.

A Supplementary Figures

The Supplementary Figures 5, 6, 7, 8, and 9, of the paper II are in included below.

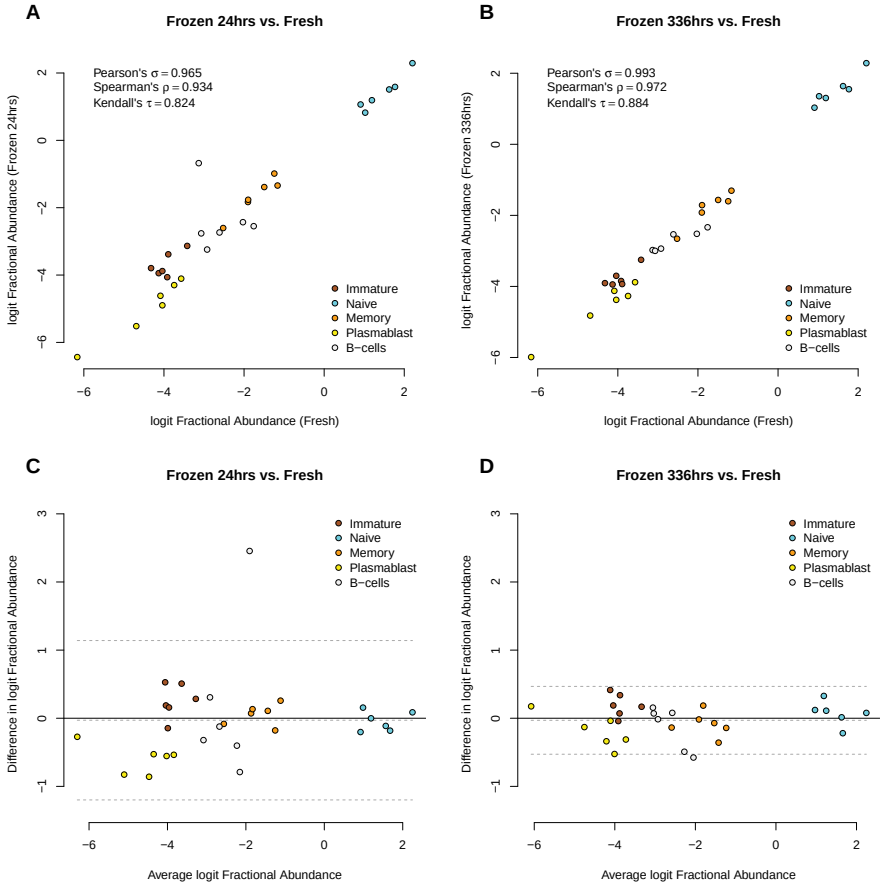


Fig. 5: Correlation and agreement of subset frequencies. The figures illustrate the correlation of the frequencies of the defined subsets for cryopreserved and fresh peripheral blood samples assessed by Pearson's, Spearman's, and Kendall's correlation coefficients. Panels A and B illustrate fresh versus 24 hours or 336 hours cryopreserved sample data for all subsets, respectively. Panel C and D illustrates the corresponding Bland-Altman plots with 95% limits of agreement and arithmetic mean difference as dashed lines.

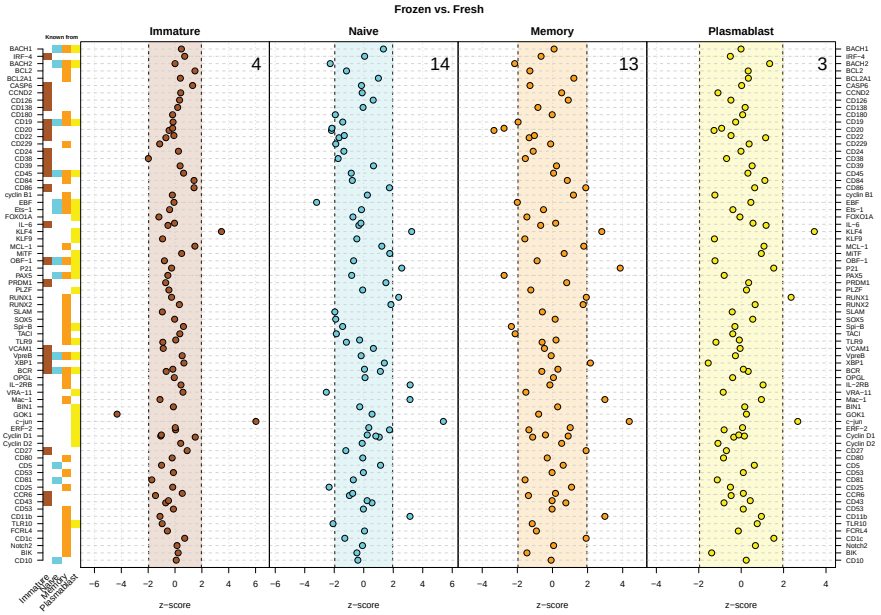


Fig. 6: Differential expression of B-cell specific differentiation and transcription markers Illustration of the analysis by the LMM z-scores for differential expression across the 4 peripheral blood B-cell subsets comparing cryopreserved and fresh samples ($N = 3$) for literature identified genes including CD markers and transcription factors. The list of genes expected to be expressed in the sorted B-cell subsets Im, N, M,and PB are shown on the left. Note, the stippled lines, showing the 95% acceptance regions, are not adjusted for multiple testing. The numbers in the upper right of each panel show the number of significant z-scores, i.e. genes outside the acceptance region.

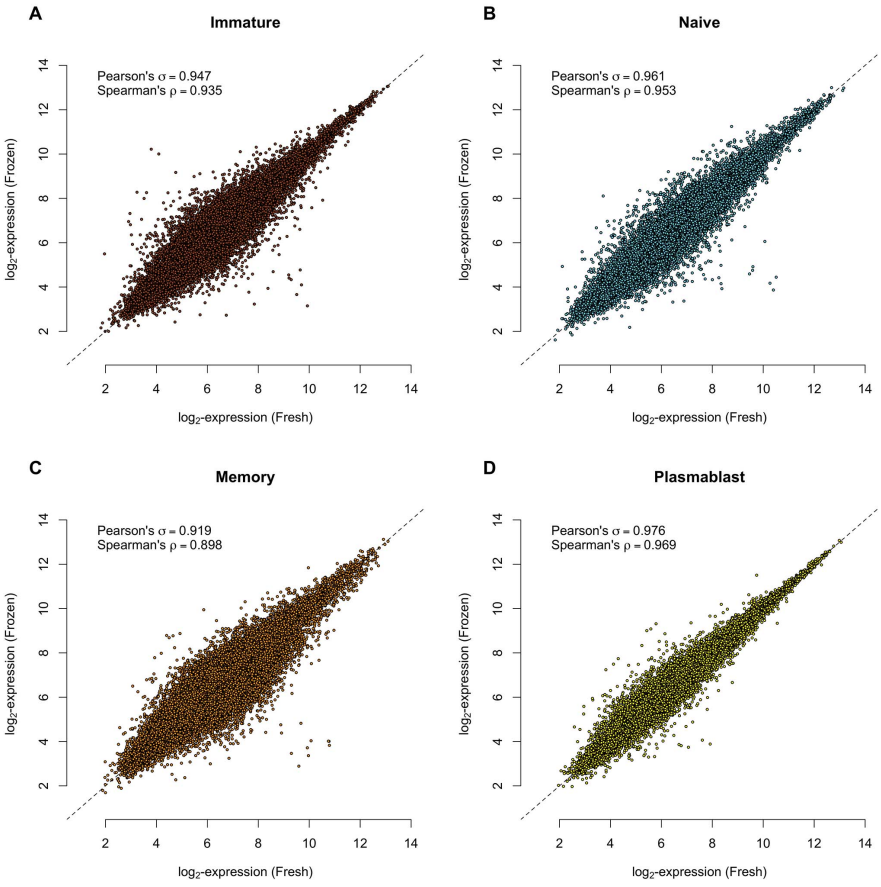


Fig. 7: Correlation of gene expression in B-cell subsets. Scatter plots of the $\log_2$-transformed probe set intensities for cryopreserved versus fresh peripheral blood samples for (A) immature, (B) memory, (C) naïve, and (D) plasmablast FACS sorted cells. The Pearson's and Spearman's correlations coefficients are also shown.

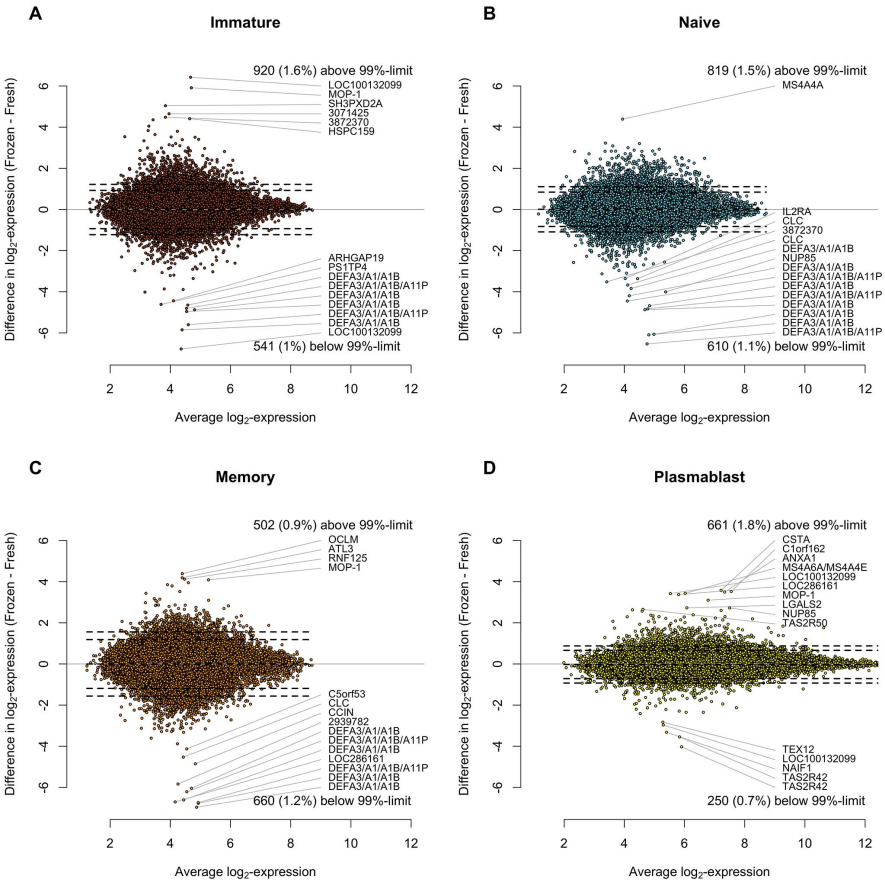


Fig. 8: Agreement of gene expression in B-cell subsets. Bland-Altman plots for each subpopulation corresponding to Supplementary Figure 7 with 95% and 99% limits of agreement and the arithmetic mean of the difference in $\log_2$ expression visualized as dashed lines. A solid line is drawn for the value of zero. This figure is essentially Figure 7 rotated 45 degrees. The number (and percentage) of genes above and below the 99% limits of agreement are shown. Furthermore, the 15 largest differences are annotated with the gene symbol.

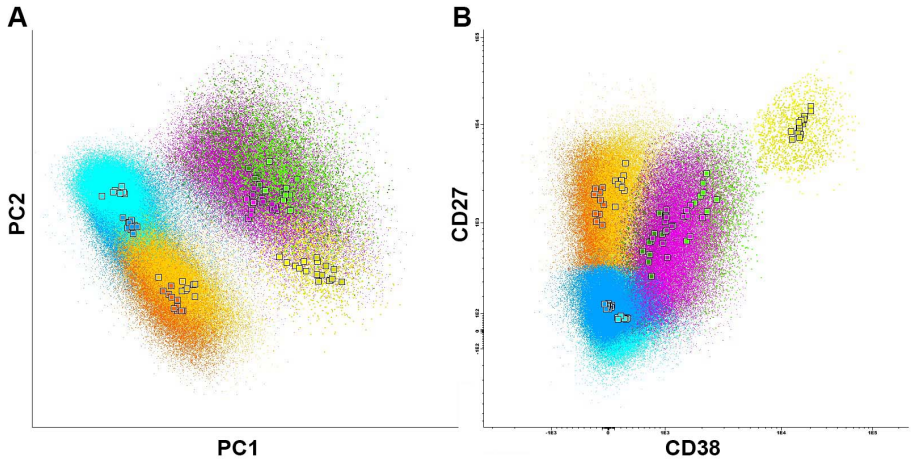


Fig. 9: Comparison of immunophenotyping of consecutive but unpaired tonsil tissue samples. Illustration of A: PCA results from a MFC multi-tube summary file merged from individual data files ($N = 18$) from fresh and cryopreserved tissue samples and B: MFC multitube summary file clusters displayed in the classical CD27 CD38 plot view. Classical B-cell subset clusters are color coded as in Figure **??** and the squares show the mean value for the software gated B-cell subsets.

Paper III

Estimation of a Common Covariance Matrix for Multiple Classes with Applications in Meta- and Discriminant Analysis

Anders Ellern Bilgrau, Poul Svante Eriksen, Karen Dybkær, and Martin Bøgsted

Submitted to
Annals of Applied Statistics, 2015.

Preface: This paper was a first attempt in modeling the covariance matrices (or their inverses) of multiple groups or studies. The aim was a tool for aggregating the covariance information across multiple studies. As such, it provides a more sophisticated alternative to the pooled covariance (and its inverse).

The model is a hierarchical model inspired by traditional random effects models and the paper explores its properties and estimation is presented. The two latent parameters of the model control the population covariance matrices of each group and their similarity. As it is a hierarchical model, it utilizes the EM algorithm to estimate the parameters.

While the model is relatively simple much of the mathematics quickly becomes cumbersome. However, we notably arrived at a very simple expression for an intra-class correlation coefficient analogue.

An obvious point of dispute is the model's ability to explain the inter-group variation by a single parameter. The model is strictly low-dimensional in the sense that the total number of samples need be larger than the dimension. The premise of the application, however, was that a sufficient number of samples could be gathered from online databases.

The paper can also be found at

`http://arxiv.org/abs/1503.07990`

and the statistical implementation in C++ and R is available at

`https://github.com/AEBilgrau/correlateR`

in the currently unfinished R-package **correlateR**. The **DLBCLdata** package (Package II) was also used in this work.

 The layout has been revised.

Estimation of a Common Covariance Matrix for Multiple Classes with Applications in Meta- and Discriminant Analysis

Abstract

We propose a hierarchical random effects model for a common covariance matrix in cases where multiple classes are present. It is applicable where the classes are believed to share a common covariance matrix of interest obscured by class-dependent noise. As such, it provides a basis for integrative or meta-analysis of covariance matrices where the classes are formed by datasets. Our approach is inspired by traditional meta-analysis using random effects models but the model is also shown to be applicable as an intermediate between linear and quadratic discriminant analysis. We derive basic properties and estimators of the model and compare their properties. Simple inference and interpretation of the introduced parameter measuring the inter-class homogeneity is suggested.

1 Introduction

The fundamental problem in statistics of accurately and precisely estimating the covariance matrix (or its inverse) is notoriously difficult. The usual bias-corrected maximum likelihood estimator (MLE), the sample covariance matrix, has long been known to perform poorly in general due to high variability [6]. The sample covariance becomes increasingly ill-conditioned as the number of variables p approaches the sample size n and singular when p exceeds n. Because of its central statistical role the list of statistical methods and applications utilizing the estimated covariance matrix is exceedingly long. Beside the many standard statistical methods such as principal component analysis (PCA), linear discriminant analysis (LDA), and quadratic discriminant analysis (QDA), examples of direct applications include gene and protein network analysis [2], spectroscopic imaging [22], functional magnetic resonance imaging (fMRI), financial forecasting, and many more. Among this expanding list of applications is also an increasing number of high-dimensional applications and datasets publicly available at online repositories.

In high-dimensional datasets the number of features p often far exceed the number of samples n. Since the number of parameters increases quadratically in p and the sample covariance matrix becomes singular when $p > n$ a plethora of shrinkage and regularization estimators have been proposed to combat the accompanying problems by effectively increasing the degrees of freedom. These examples include the graphical LASSO and ridge estimation of the precision matrix [14, 27]. Instead of attempting to derive still more sophisticated estimators we attempt to alleviate the problem from a different angle by using more available data and thus effectively increasing n. While the high-dimensional

extension to $p > n$ is important it is out of scope in this paper. We restrict ourselves to the case where the total number of samples exceed p. Hence, if k classes or datasets are available with sample sizes $n_1, \ldots, n_k$, we consider the case where $p < \sum_{i=1}^{k} n_i$ while allowing p to exceed n_i for each individual class i.

As with all major groups of cancer, a large number of diffuse large B-cell lymphoma (DLBCL) genomic datasets are now publicly available online. We wanted to use these studies in combination with data from our own laboratory to arrive at a good estimate of the covariance matrix whilst accounting for and assessing inter-study variation. Although this work was motivated by gene-gene interaction networks in DLBCL, where the covariance matrix is assumed to contain all information about the conditional dependencies of the genes, the methods are general and not limited to such genomic data.

2 A random effects model for the covariance matrix

The model below was motivated by ordinary meta-analysis. Meta-analysis comes in various flavors corresponding to the assumption on the nature of the inter-study treatment effect. Random-effects models (REM) in meta-analysis model the inter-study effects as random variables [3, 8]. In a vein similar to the ordinary meta-analysis approach, we think of the different studies as related but perturbed experiments and propose the following simple random covariance model (RCM) of the observations. Let p be the number of features and k the number of classes. We model an observation $\mathbf{x}$ from the i'th study as a p-dimensional zero-mean multivariate gaussian vector with covariance matrix realized from an inverse Wishart distribution, i.e. $\mathbf{x}$ follows the hierarchical model

$$\begin{aligned} \boldsymbol{\Sigma}_i &\sim \mathcal{W}_p^{-1}\big((\nu - p - 1)\boldsymbol{\Sigma}, \nu\big), \\ \mathbf{x}|\boldsymbol{\Sigma}_i &\sim \mathcal{N}_p(\mathbf{0}_p, \boldsymbol{\Sigma}_i), \qquad i = 1, ..., k, \end{aligned} \tag{1}$$

where $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_i)$ denotes a p-dimensional multivariate gaussian distribution with mean $\boldsymbol{\mu}$, positive definite (p.d.) covariance matrix $\boldsymbol{\Sigma}_i$, and probability density function (pdf)

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_i) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}_i|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu})\right).$$

As seen, we use the generic notation $f(\cdot|\cdot)$ and $f(\cdot)$ for the conditional and unconditional pdf of random variables, respectively, throughout this paper. Above, $\mathcal{W}_p^{-1}(\boldsymbol{\Psi}, \nu)$ denotes a p-dimensional inverse Wishart distribution with ν degrees of freedom, a p.d. $p \times p$ scale matrix $\boldsymbol{\Psi}$, and pdf

$$f(\boldsymbol{\Sigma}_i) = \frac{|\boldsymbol{\Psi}|^{\frac{\nu}{2}}}{2^{\frac{\nu p}{2}} \Gamma_p\big(\frac{\nu}{2}\big)} |\boldsymbol{\Sigma}_i|^{-\frac{\nu+p+1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\big(\boldsymbol{\Psi}\boldsymbol{\Sigma}_i^{-1}\big)\right), \quad \nu > p - 1 \tag{2}$$

where $\mathbf{\Sigma}_i$ is p.d. and Γ_p is the multivariate generalization of the gamma function Γ seen in Appendix B.1. While the inverse Wishart distribution is defined for all $\nu > p-1$, the first order moment $(\nu - p - 1)^{-1}\mathbf{\Psi}$ exists only when $\nu > p+1$. With the reparameterization $\mathbf{\Psi} = (\nu - p - 1)\mathbf{\Sigma}$ in the inverse Wishart distribution of (1) the common expected covariance matrix is

$$\mathbf{\Sigma} = \mathbb{E}[\mathbf{\Sigma}_i] = \frac{\mathbf{\Psi}}{\nu - p - 1} \text{ for } \nu > p + 1. \tag{3}$$

Hence in the RCM given by (1), $\mathbf{\Sigma}$ can be interpreted as a location-like parameter as it is the expected covariance matrix in each study. The parameter ν inversely controls the inter-class variation and can thus be considered an inter-class homogeneity parameter of the covariance structure. A large ν corresponds to high study homogeneity and vice versa for small ν. This can be further seen as $\mathbf{\Sigma}_i$ concentrates around $\mathbf{\Sigma}$ for $\nu \to \infty$ which can be interpreted as the inter-study variation goes towards zero for increasing ν. Thus, the true underlying covariance matrix $\mathbf{\Sigma}$ and the homogeneity parameter ν are the effects of interest to be estimated in this paper.

These basic properties of the RCM motivates the construction. We note that while the reparameterization of (1) has a preferable interpretation, the likelihood is much more complex and often numerically unstable. The reparameterization is especially problematic for ν near $p+1$ and indeed senseless when the expected covariance cease to exist for $p-1 < \nu \leq p+1$. Therefore, we use the usual parameterization by $\mathbf{\Psi}$ in the fitting procedure and the remainder of this paper.

2.1 The likelihood function

Suppose $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i}$ are n_i i.i.d. observations from $i = 1, ..., k$ independent studies from the model given in (1). Let $\mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i})^\top$ be the $n_i \times p$ matrix of observations for the i'th study where rows correspond to samples and columns to variables. By the independence assumptions, the log-likelihood for $\mathbf{\Psi}$ and ν is given by

$$\begin{aligned}
&\ell\big(\mathbf{\Psi}, \nu \big| \mathbf{X}_1, ..., \mathbf{X}_k\big) = \log f\big(\mathbf{X}_1, ..., \mathbf{X}_k \big| \mathbf{\Psi}, \nu\big) \\
&= \log \int f(\mathbf{X}_1, ..., \mathbf{X}_k | \mathbf{\Sigma}_1, ..., \mathbf{\Sigma}_k, \mathbf{\Psi}, \nu) f(\mathbf{\Sigma}_1, ..., \mathbf{\Sigma}_k | \mathbf{\Psi}, \nu) d\mathbf{\Sigma}_1 \cdots d\mathbf{\Sigma}_k \\
&= \log \prod_{i=1}^{k} \int f(\mathbf{X}_i | \mathbf{\Sigma}_i) f(\mathbf{\Sigma}_i | \mathbf{\Psi}, \nu) d\mathbf{\Sigma}_i.
\end{aligned}$$

Since the inverse Wishart distribution is conjugate to the multivariate gaussian distribution the integral, of which the integrand forms a gaussian-inverse-Wishart distribution, can be evaluated. Hence $\mathbf{\Sigma}_i$ can be marginalized out, cf. (13) in Appendix A, and we arrive at the following expression for the log-

likelihood function,

$$\ell\big(\boldsymbol{\Psi},\nu\big|\mathbf{X}_1,...,\mathbf{X}_k\big) = \log\prod_{i=1}^{k}\frac{\big|\boldsymbol{\Psi}\big|^{\frac{\nu}{2}}\Gamma_p\big(\frac{\nu+n_i}{2}\big)}{\pi^{\frac{n_i p}{2}}\big|\boldsymbol{\Psi}+\mathbf{X}_i^\top\mathbf{X}_i\big|^{\frac{\nu+n_i}{2}}\Gamma_p\big(\frac{\nu}{2}\big)}$$
$$= \sum_{i=1}^{k}\bigg[\frac{\nu}{2}\log\big|\boldsymbol{\Psi}\big| - \frac{\nu+n_i}{2}\log\big|\boldsymbol{\Psi}+\mathbf{X}_i^\top\mathbf{X}_i\big| - \log\frac{\Gamma_p\big(\frac{\nu+n_i}{2}\big)}{\Gamma_p\big(\frac{\nu}{2}\big)}\bigg], \tag{4}$$

up to an additive constant. As should be expected, the scatter matrix $\mathbf{S}_i = \mathbf{X}_i^\top\mathbf{X}_i$ and study sample size n_i are sufficient statistics for each study. Note that $\mathbf{S}_i$ is conditionally Wishart distributed, $\mathbf{S}_i|\boldsymbol{\Sigma}_i \sim \mathcal{W}(\boldsymbol{\Sigma}_i, n_i)$, by construction.

As stated in the following two propositions the likelihood is not log-concave in general. However, it is log-concave as a function of ν.

Proposition 1 (Non-concavity in $\boldsymbol{\Psi}$)
For fixed ν, the log-likelihood function (4) is not concave in $\boldsymbol{\Psi}$.

All proofs have been deferred to Appendix B.

Proposition 2 (Concavity in ν)
For fixed positive definite $\boldsymbol{\Psi}$, the log-likelihood function (4) is concave in ν.

While the likelihood is not concave in $\boldsymbol{\Psi}$ we are able to show the existence and uniqueness of a global maximum in $\boldsymbol{\Psi}$.

Proposition 3 (Existence and uniqueness)
The log-likelihood (4) has a unique maximum in $\boldsymbol{\Psi}$ for fixed ν and $n_\bullet = \sum_{a=1}^{k} n_a \geq p$.

This result is proven in Appendix B and follows from two lemmas stated therein.

In the following section estimators of the parameters are derived using moments and the EM algorithm assuming ν fixed.

2.2 Moment estimator

The pooled empirical covariance matrix can be viewed as a moment estimator of $\boldsymbol{\Sigma}$. By the assumptions the first and second moment of the j'th observation in the i'th study, $\mathbf{x}_{ij}$, is given by $\mathbb{E}[\mathbf{x}_{ij}] = \mathbf{0}_p$ and

$$\mathbb{E}[\mathbf{x}_{ij}\mathbf{x}_{ij}^\top] = \mathbb{E}\big[\mathbb{E}[\mathbf{x}_{ij}\mathbf{x}_{ij}^\top|\boldsymbol{\Sigma}_i]\big] = \mathbb{E}[\boldsymbol{\Sigma}_i] = \frac{\boldsymbol{\Psi}}{\nu-p-1} = \boldsymbol{\Sigma}.$$

for all $j = 1, ..., n_i$ and $i = 1, ..., k$. This suggests the estimators

$$\hat{\mathbf{\Psi}}_{\text{pool}} = (\nu - p - 1)\frac{\sum_{i=1}^{k} \mathbf{S}_i}{\sum_{i=1}^{k} n_i} \text{ and } \hat{\mathbf{\Sigma}}_{\text{pool}} = \frac{\sum_{i=1}^{k} \mathbf{S}_i}{\sum_{i=1}^{k} n_i}, \qquad \nu > p + 1 \tag{5}$$

where the latter is obtained by plugging $\hat{\mathbf{\Psi}}_{\text{pool}}$ into (3). This is the well-known pooled empirical covariance matrix.

2.3 Maximization using the EM algorithm

Here the updating scheme of the expectation-maximization (EM) algorithm [7] for fixed ν is derived. We now compute the expectation step of the EM-algorithm.

From (1) we have that,

$$\begin{aligned} \mathbf{\Sigma}_i &\sim \mathcal{W}_p^{-1}(\mathbf{\Psi}, \nu), \\ \mathbf{S}_i|\mathbf{\Sigma}_i &\sim \mathcal{W}_p(\mathbf{\Sigma}_i, n_i) \quad \text{for } i = 1, ..., k. \end{aligned}$$

Let $\mathbf{\Delta}_i = \mathbf{\Sigma}_i^{-1}$ be the precision matrix and let $\mathbf{\Theta} = \mathbf{\Psi}^{-1}$, then we equivalently have that

$$\begin{aligned} \mathbf{\Delta}_i &\sim \mathcal{W}_p(\mathbf{\Theta}, \nu), \\ \mathbf{S}_i|\mathbf{\Delta}_i &\sim \mathcal{W}_p(\mathbf{\Delta}_i^{-1}, n_i). \end{aligned} \tag{6}$$

From the conjugacy of the inverse Wishart and the Wishart distribution, the posterior distribution of the precision matrix is

$$\mathbf{\Delta}_i|\mathbf{S}_i \sim \mathcal{W}_p\Big(\big(\mathbf{\Theta}^{-1} + \mathbf{S}_i\big)^{-1}, n_i + \nu\Big).$$

Hence, by the expectation of the Wishart distribution,

$$\mathbb{E}[\mathbf{\Delta}_i|\mathbf{S}_i] = (n_i + \nu)\big(\mathbf{\Theta}^{-1} + \mathbf{S}_i\big)^{-1}.$$

The maximization step, in which the log-likelihood $\ell(\mathbf{\Theta}|\mathbf{\Delta}_1, ..., \mathbf{\Delta}_k)$ is maximized, yields the estimate $\hat{\mathbf{\Theta}} = \frac{1}{k\nu}\sum_{i=1}^{k} \mathbf{\Delta}_i$, which is the mean of the scaled precision matrices $\frac{1}{\nu}\mathbf{\Delta}_i$. The derivation of this estimate can be seen in Appendix C. Let $\hat{\mathbf{\Theta}}_{(t)}$ be the current estimate of $\mathbf{\Theta}$. This yields the updating scheme

$$\hat{\mathbf{\Theta}}_{(t+1)} = \frac{1}{k\nu}\sum_{i=1}^{k}(n_i + \nu)\left(\hat{\mathbf{\Theta}}_{(t)}^{-1} + \mathbf{S}_i\right)^{-1} \tag{7}$$

for $\mathbf{\Theta}_{(t)}$. We denote the inverse of the estimate obtained by repeated iteration of (7) by $\hat{\mathbf{\Psi}}_{\text{EM}}$.

An approximate maximum likelihood estimator using a first order approximation is also possible. This derivation has been deferred to Appendix D.

Algorithm 1 RCM coordinate ascent estimation procedure

Input:
Sufficient data: $(\mathbf{S}_1, n_1), ..., (\mathbf{S}_k, n_k)$
Initial parameters: $\hat{\mathbf{\Psi}}_{(0)}, \hat{\nu}_{(0)}$
Convergence criterion: $\varepsilon > 0$
Output:
Parameter estimates: $\hat{\mathbf{\Psi}}, \hat{\nu}$
procedure FITRCM($\mathbf{S}_1, ..., \mathbf{S}_k, n_1, ..., n_k, \hat{\mathbf{\Psi}}_{(0)}, \hat{\nu}_{(0)}, \varepsilon$)
    Initialize: $l_{(0)} \leftarrow \ell(\hat{\mathbf{\Psi}}_{(0)}, \hat{\nu}_{(0)})$
    for $t = 1, 2, 3, ...$ **do**
        $\hat{\mathbf{\Psi}}_{(t)} \leftarrow U\left(\hat{\mathbf{\Psi}}_{(t-1)}, \hat{\nu}_{(t-1)}\right)$
        $\hat{\nu}_{(t)} \leftarrow \arg\max_{\nu} \ell\left(\hat{\mathbf{\Psi}}_{(t)}, \nu\right)$
        $l_{(t)} \leftarrow \ell\left(\hat{\mathbf{\Psi}}_{(t)}, \hat{\nu}_{(t)}\right)$
        if $l_{(t)} - l_{(t-1)} < \varepsilon$ **then**
            return $\left(\hat{\mathbf{\Psi}}_{(t)}, \nu_{(t)}\right)$
        end if
    end for
end procedure

2.4 Estimation procedure

We propose a procedure alternating between estimating ν and $\mathbf{\Psi}$ while keeping the other fixed. Given parameters $\hat{\nu}_{(t)}$ and $\hat{\mathbf{\Psi}}_{(t)}$ at iteration t, we estimate $\hat{\mathbf{\Psi}}_{(t+1)}$ using fixed $\hat{\nu}_{(t)}$. Subsequently, we find $\hat{\nu}_{(t+1)}$ by a standard one-dimensional numerical optimization procedure using the fixed $\hat{\mathbf{\Psi}}_{(t+1)}$. This coordinate ascent approach is repeated until convergence as described in Algorithm 1. The update function U in the algorithm is defined by the derived estimators. That is, equations (5), (7), or (21) define U to be defined by the pooled, EM, or approximate MLE estimates, respectively.

The procedure using the EM step utilizes the results about the RCM log-likelihood and thus provides a guarantee of convergence along with the advantage of a very simple implementation. Both the EM step and the ν update will always yield an increase in the likelihood. However, the obvious disadvantage is that the identified maxima might be a saddle-point when considering the log-likelihood function jointly in $(\mathbf{\Psi}, \nu)$.

2.5 Interpretation and inference

Test for no class heterogeneity

By the RCM construction ν parameterizes an inter-class variance where the size of ν corresponds to the homogeneity between the classes. A large ν yields

high study homogeneity while small ν yields low homogeneity. Thus it might be of interest to test if the estimated homogeneity $\hat{\nu}$ is extreme under the null-hypothesis of no heterogeneity (i.e. infinite homogeneity). I.e. a test for the hypothesis $H_0 : \nu = \infty$ which is equivalent to

$$H_0 : \mathbf{\Sigma}_1 = ... = \mathbf{\Sigma}_k = \mathbf{\Sigma}.$$

The two are equivalent since sampling the covariance matrix from the inverse Wishart distribution becomes deterministic for $\nu = \infty$. Testing this hypothesis can therefore also be interpreted as testing if the data is adequately explained when leaving out the hierarchical structure.

The distribution of $\hat{\nu}$ under the null hypothesis is not tractable. However in practice, under H_0 or when ν is extremely large the estimated $\hat{\nu}_{\text{obs}}$ will be finite as the intra-study variance dominates the total variance. We note that the null distribution of $\hat{\nu}$ does not depend on $\mathbf{\Sigma}$. We propose approximating the distribution of $\hat{\nu}$ under H_0 by resampling. To do this, the model is simply fitted a large number of times N on datasets re-sampled under H_0 mimicked by permuted class labels to get $\hat{\nu}_0^{(1)}, ..., \hat{\nu}_0^{(N)}$. As *small* values of $\hat{\nu}$ are critical for H_0 approximate acceptance regions can be constructed from $\hat{\nu}_0^{(j)}, j = 1, ..., N$. Likewise, an approximation of the p value testing H_0 can be obtained by

$$P = \frac{1}{N+1}\left(1 + \sum_{j=1}^{N} \mathbb{1}\left[\hat{\nu}_0^{(j)} < \hat{\nu}_{\text{obs}}\right],\right)$$

where $\mathbb{1}[\,\cdot\,]$ is the indicator function. The addition of one in the nominator and denominator adds a positive bias to the approximate p-value minimally needed according to Phipson and Smyth [24]. This is approximately the fraction of $\hat{\nu}_0^{(j)}$'s smaller than $\hat{\nu}_{\text{obs}}$.

Intra-class correlation coefficient

We now introduce a descriptive statistic analogous to the intra-class correlation coefficient (ICC) [26] well known from ordinary meta-analysis to better determine what constitute large values of ν. For the RCM, the ICC be given by

$$\text{ICC}(\nu) = \frac{1}{\nu - p}. \tag{8}$$

This follows from the definition of the ICC which is the ratio of the between-study variation (Σ_{ij}) and the total variation (S_{ij}) of a single pair of any variables. Consider observations from (1). We temporarily abuse our notation and let

$$\mathbf{\Sigma} \sim \mathcal{W}_p^{-1}(\mathbf{\Psi}, \nu) \quad \text{and} \quad \mathbf{S}|\mathbf{\Sigma} \sim \mathcal{W}_p(\mathbf{\Sigma}, 1),$$

and consider only a single observation ($n = 1$). Furthermore, let $\mathbf{S} = (S_{ij})_{p\times p}$, $\mathbf{\Sigma} = (\Sigma_{ij})_{p\times p}$, and $\mathbf{\Psi} = (\Psi_{ij})_{p\times p}$. To compute the ICC, we are thus interested in the ratio of the quantities $\mathbb{V}(\Sigma_{ij})$ and $\mathbb{V}(S_{ij})$ corresponding to the between-study and total variation of the covariance between variables i and j, respectively. That is, the ICC is the proportion of the total variance between studies,

$$\mathrm{ICC}(\nu) = \frac{\mathbb{V}(\Sigma_{ij})}{\mathbb{V}(S_{ij})} = \frac{\mathbb{V}(\Sigma_{ij})}{\mathbb{V}(\Sigma_{ij}) + \mathbb{E}[\mathbb{V}(S_{ij}|\mathbf{\Sigma})]}, \tag{9}$$

where the second equality is obtained by $\mathbb{E}[S_{ij}|\mathbf{\Sigma}] = \Sigma_{ij}$ and the law of total variation. This equality agrees with the usual ICC as $\mathbb{E}[\mathbb{V}(S_{ij}|\Sigma_{ij})]$ can be interpreted as the (expected) within-study variation. Using the conditional variance given by $\mathbb{V}(S_{ij}|\mathbf{\Sigma}) = \Sigma_{ij}^2 + \Sigma_{ii}\Sigma_{jj}$ the needed quantities can be found. To compute an expression for (9) we need to consider the fourth-order moments of the observations. From the model, known results of the inverse Wishart distribution, cf. [4, 28], leads to

$$\mathrm{Cov}(\Sigma_{ij}, \Sigma_{kl}) = \frac{2\Psi_{ij}\Psi_{kl} + (\nu-p-1)\big(\Psi_{ik}\Psi_{jl} + \Psi_{il}\Psi_{kj}\big)}{(\nu-p)(\nu-p-1)^2(\nu-p-3)},\ \nu > p+3, \tag{10}$$

implying that

$$\mathbb{V}(\Sigma_{ij}) = \mathrm{Cov}(\Sigma_{ij}, \Sigma_{ij}) = \frac{(\nu-p+1)\Psi_{ij}^2 + (\nu-p-1)\Psi_{ii}\Psi_{jj}}{(\nu-p)(\nu-p-1)^2(\nu-p-3)}. \tag{11}$$

We continue with the conditional variance of $S_{ij}|\mathbf{\Sigma}$ in the denominator of (9),

$$\begin{aligned}\mathbb{E}\big[\mathbb{V}(S_{ij}|\Sigma_{ij})\big] &= \mathbb{V}(\Sigma_{ij}) + \mathbb{E}[\Sigma_{ij}]^2 + \mathrm{Cov}(\Sigma_{ii}, \Sigma_{jj}) + \mathbb{E}[\Sigma_{ii}]\mathbb{E}[\Sigma_{jj}] \\ &= \mathbb{V}(\Sigma_{ij}) + \mathrm{Cov}(\Sigma_{ii}, \Sigma_{jj}) + (\nu-p-1)^{-2}(\Psi_{ij}^2 + \Psi_{ii}\Psi_{jj}).\end{aligned} \tag{12}$$

An expression of $\mathbb{V}(S_{ij})$ in terms of the elements of $\mathbf{\Psi}$ can then found by substituting (10) and (11) into (12) and by extension an expression for the ICC (9) can be obtained. We omit this tedious calculation which can be verified to yield $\mathrm{ICC}(\nu) = 1/(\nu-p)$ above. Naturally enough, the ICC depends only on ν. A straight-forward plug-in estimator $\widehat{\mathrm{ICC}}(\nu)$ of the ICC of some gene-gene interaction is then $\mathrm{ICC}(\hat{\nu})$.

Though $v > p+3$ is required for the variances to exist, it is clear that $\mathrm{ICC}(\nu) \to 1$ for $\nu \to (p+1)^+$ and $\mathrm{ICC}(\nu) \to 0$ for $\nu \to \infty$ as should be expected.

3 Assessment of the estimation procedures

To assess the precision and stability of the estimation procedure we generated data from the hierarchical model (1) for $p = 10$ variables and $k = 3$ studies each with an equal number of observations, $n = n_1 = n_2 = n_3$. We chose the

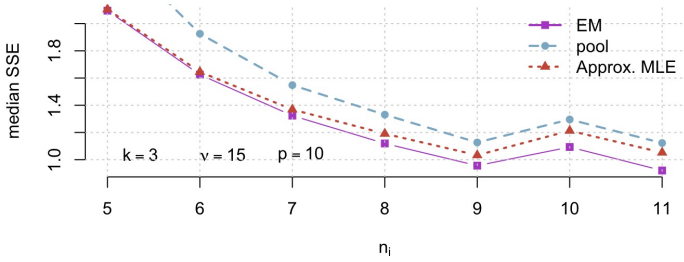


Fig. 1: The median SSE of 2500 simulations as a function of the number of samples n_i in each of the 3 studies.

parameters $\nu = 15$ and $\Psi_{ij} = 1$ for all $i = j$ and $\Psi_{ij} = 0.5$ for all $i \neq j$. The number of observations in each study n was varied in the range $[5, 11]$.

We measured the precision of the estimated values against the expected covariance matrix given by (3). Let $\hat{\mathbf{\Psi}}$ and $\hat{\nu}$ be the estimates obtained using the moment, EM, or approximate MLE (defined in Appendix D) approaches as described. We benchmark the proposed estimators against the known truth. The benchmarking measure used is the following weighted sum of squared errors,

$$\text{SSE}(\hat{\mathbf{\Sigma}}) = \sum_{i \leq j} \frac{(\hat{\Sigma}_{ij} - \Sigma_{ij})^2}{(\Sigma ij)} \quad \text{where} \quad (\Psi_{ij}) = n(\Sigma_{ij}^2 + \Sigma_{ii}\Sigma_{jj}).$$

For each $n = 5, \ldots, 11$, the weighted sum of squared errors for each estimator, $\text{SSE}(\hat{\mathbf{\Sigma}})$, were computed for 2500 datasets and the median of these values are seen in Figure 1 as function of the number of samples in each dataset n_i.

We see that the EM estimation is superior to that of the approximate MLE and moment estimators.

3.1 Implementation and availability

Algorithm 1 and the different estimators are implemented in the statistical programming language R [25] with core functions in `C++` using packages Rcpp and RcppArmadillo [11, 12]. They are incorporated in the open-source R-package `correlateR` freely available for forking and editing at `http://github.com/AEBilgrau/correlateR`. We refer to the information here for further details and installation instructions. This document was prepared with `knitr` [29] and LaTeX. To reproduce this document see 5.

4 Applications

4.1 DLBCL meta-analysis

Diffuse large B-cell lymphoma (DLBCL) is an aggressive cancer subtype accounting for $30\% - 58\%$ of all non-Hodgkin's lymphomas (NHL) which itself constitutes about 90% of all lymphomas [20].

Data and preprocessing

A large amount of DLBCL gene expression datasets are now available online at the NCBI (National Center for Biotechnology Information) Gene Expression Omnibus (GEO) website. Ten large-scale DLBCL gene expression studies were downloaded and preprocessed using custom brainarray chip definition files (CDF) [5] and RMA-normalized using the R-package `affy` [17]. The corresponding GEO-accession numbers are GSE12195, GSE22895, GSE31312, GSE10846, GSE34171, GSE22470, GSE4475, and GSE19246 based on various microarray platforms. The downloaded data together with a dataset from our own laboratory (GSE56315, [10]) yields a total of 2046 samples with study sizes in the range 78-469. The summarization using brainarray CDFs to Ensembl gene identifiers facilitates cross-platform integration.

After RMA normalization and summarization, the data were brought to a common scale by quantile normalizing all data to the common cumulative distribution function of all arrays. Lastly, the datasets were reduced to 11573 common genes represented in all studies and array platforms.

Analysis

A coexpression network (or weighted correlation network) analysis integrating all datasets was carried out. For each dataset the scatter matrix $\mathbf{S}_i$ of the top 300 most variable genes (as measured by the pooled variance across all studies) was computed as the sufficient statistics along with the number of samples. Hence, we investigate 45,150 pairwise interactions.

The parameters of the RCM were estimated using the EM algorithm and yielded the 300×300 matrix $\hat{\mathbf{\Psi}}$ and $\hat{\nu} = 773.16$. From these, $\hat{\mathbf{\Sigma}} = (\hat{\nu} - p - 1)^{-1}\hat{\mathbf{\Psi}}$ was computed and subsequently scaled to the corresponding correlation matrix $\hat{\mathbf{R}}$.

The estimated $\hat{\nu}$ yields a surprisingly low estimated ICC of 0.0021. Hence by the RCM, only 0.21% of the variability of the gene-gene covariances is between-studies on average. The selection of only the most (within study) varying genes is an obvious contribution to the low ICC. Hence, by selection we have high within-study variability. An alternative contribution could indeed also be high study homogeneity. In any case, the low ICC might suggest high reproducibility of the covariances between studies of the selected genes.

Next, we outline one of many possible downstream analysis of the estimated covariance. For simplicity we employed standard correlation network analyses to the estimated common correlation matrix $\hat{\mathbf{R}}$ across all studies. To identify

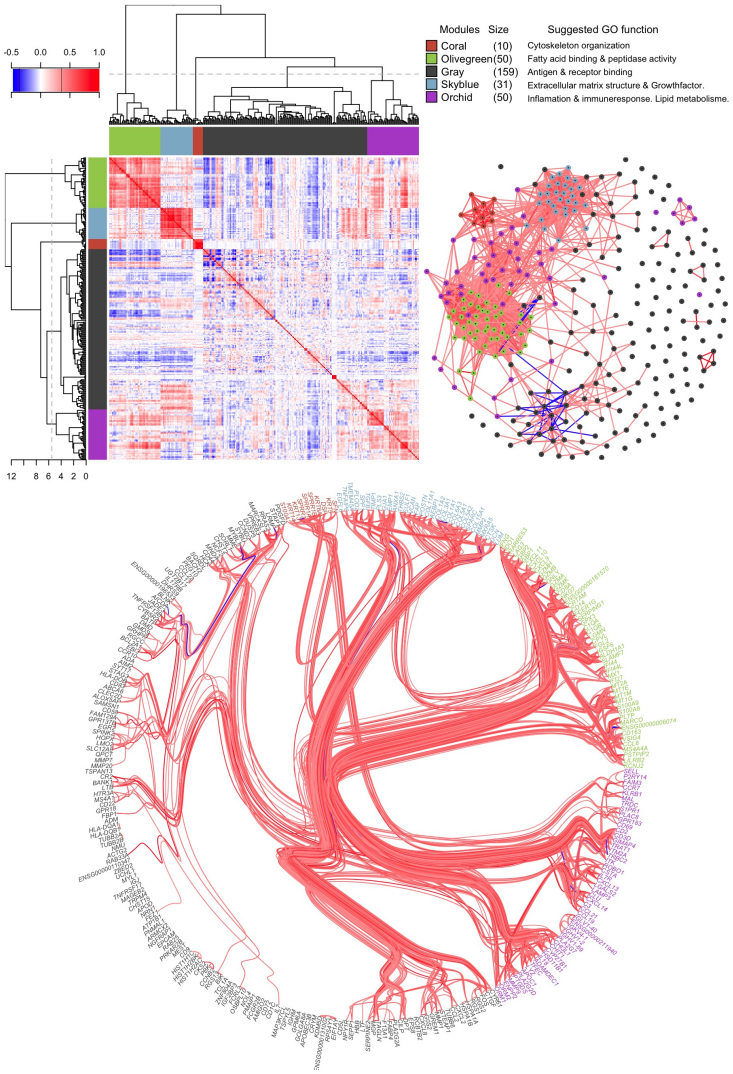


Fig. 2: Top left: dendrograms of the hierarchical clustering, the identified modules, and a heatmap of the correlation matrix. Top right: the correlation network as laid out by the Fruchterman-Reingold algorithm [16]. The nodes are colored after the identified modules. The edge colors follow the color key of the heatmap. If the edge weight is numerically less than 0.359, corresponding to the 95% largest values, the edge is suppressed. Bottom: A Hierarchical edge bundling representation of the network [18] where edges loosly follow the dendrogram. Edge colors follow the color key. Only edges with a weight numerically larger than 0.359 are plotted.

clusters with high internal correlation, we used agglomerative hierarchical clustering with Ward-linkage and distance measure defined as 1 minus the absolute value of the correlation. The tree was arbitrarily pruned at a height which produces 5 modules named by colors. Figure 2 shows these results. A heat-map, the hierarchical tree, and the identified modules are seen at the top left. The top-right shows a graphical representation of the matrix which better illustrates the clusters and their relations. The bottom plot shows the graph radially laid out using hierarchical edge bundling [18] where the edges are guided by the hierarchical tree. Table 1 shows the top genes within each module. As seen e.g. in the olivegreen module, genes from the same gene family are clustered together.

To further investigate the low ICC, we refitted the RCM on the subset of the 50 genes in the orchid module for reasons clear later. This yielded an ICC(182.4) = 0.008. Subsequently the model was repeatedly fitted on 50 randomly chosen genes 500 times to gauge the size of the ICC without the selection bias. This resulted in a mean ICC (lower quartile, upper quartile) of 0.019 $(0.018, 0.021)$ suggesting that there is a high study homogeneity under randomly selected genes.

The test for the null hypothesis of no study heterogeneity, $\nu = \infty$, is clearly rejected with a p value of 0.002. The mean (sd) of the fitted $\hat{\nu}$ on 500 permuted datasets was 2156.2 (1.39) compared to $\hat{\nu} = 182.4$ in the observed dataset.

Next, the modules were screened for biological relevance using GO (Gene Ontology) enrichment analysis. The upper right of Figure 2 shows suggested functions of the modules primarily based on the GO analysis. 5 shows the significant GO-terms at significance level 0.01 for each module in which the most GO-terms appear highly relevant to the pathology of DLBCL.

Lastly, we checked if the identified modules were prognostic for overall survival (OS) in the CHOP and R-CHOP-treated cohorts of the GSE10846 datasets. To do this, the eigengene [19] for each module was computed and a multiple Cox proportional hazards model for OS was fitted with the module eigengenes as covariates. The module eigengene is simply the first principal of the expression matrix of the module which can thus be represented by a linear combination of the module genes. For the prognostic interesting orchid module, the Kaplan-Meier estimates were computed for groups arising when dichotomizing the values of the corresponding eigengene as above or below the median value. These results are seen in Figure 3.

From the survival analysis, the orchid module appeared particularly interesting since it marked a gene cluster identifying DLBCL patients with significantly improved outcome. Therefore a manual screening of the 50 genes within the orchid module was performed. The genes CHI3L1, CHIT1, and LYZ are related to chitin degradation and suggests activated immune system response and inflammation. Enzymes related to chitin degradation can possibly originate from macrophages as CHIT1 is expressed by activated macrophages. The inflammation and modulated activity of the immune system are further suggested by the genes ORM1, PLA2G2D, PLA2G7, and IL18. CHIT3L1

Table 1: The identified modules, their sizes, and member genes. The genes are sorted decreasingly by their intra-module connectivity (sum of the incident edge weights). Only the top 20 genes are shown.

Gray	**Olivegreen**	**Orchid**	**Skyblue**	**Coral**
n = 159	n = 50	n = 50	n = 31	n = 10
RGS13	S100A8	CXCL13	COL3A1	KRT6A
BCL2A1	C1QB	CCL19	COL1A2	KRT14
MMP1	CCL8	CHI3L1	COL5A2	KRT13
GMDS	VSIG4	CLU	MXRA5	SPRR3
FEZ1	CXCL9	CCL21	SULF1	SPRR1B
HLA-DQA1	C1QA	PTGDS	POSTN	SPRR1A
IGHM	CXCL10	CD3D	THBS2	S100A2
CR2	GBP1	PLAC8	MMP2	KRT5
CD83	CXCL11	CD2	COL6A3	DSP
AICDA	GZMB	CXCL14	VCAN	SFN
HLA-DQB1	IDO1	TRAT1	LUM	
UGT2B17	MT1G	TRBC2	SPP1	
BIK	GZMA	ADAMDEC1	COL5A1	
MS4A1	GZMK	CSTA	PLOD2	
GRHPR	ALDH1A1	ITK	COL15A1	
CYB5R2	S100A9	IL7R	DCN	
RPS4Y1	FCER1G	CHIT1	CTSK	
ADA	PSTPIP2	GIMAP4	COL11A1	
DMD	LILRB2	ENPP2	COL1A1	
ACTG2	GZMH	LGALS2	FAP	

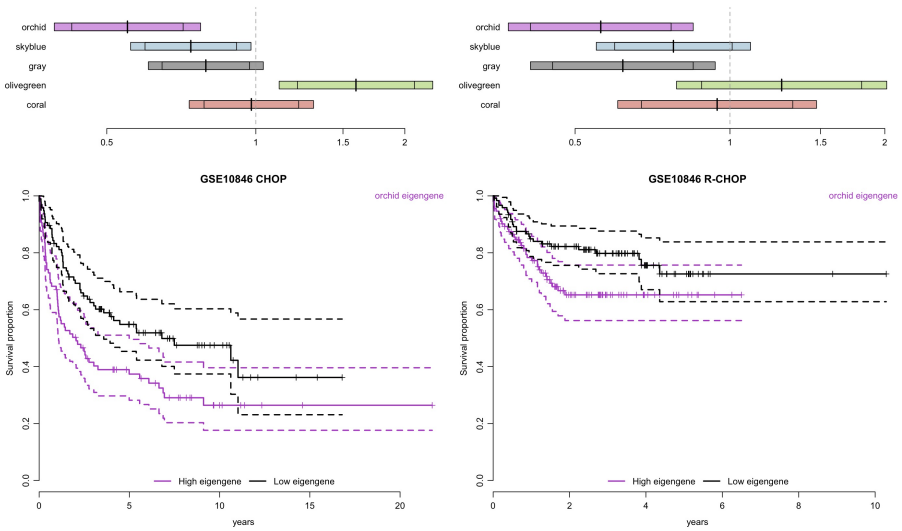


Fig. 3: The top row shows 95% and 99% CI for the hazard ratio for each eigengene in the multiple Cox proportional hazards model containing all eigengenes. The bottom row shows Kaplan-Meier estimates (and 95% CI) of the overall survival for patients stratified by the dichotomized orchid eigengene.

(also known as YKL40) has been linked to the AKT anti-apoptotic signaling pathway in glioblastoma [13] and thus high YKL40 is associated with poor outcome, Some of the remaining, MMP9, PTGDS, ADAMDEC1, HSD11B1, APOC1, and CYP27B1 are involved in metalloproteinase degradation and lipid metabolism. MMP9 in particular is known to have a central role in proliferation, migration, differentiation, angiogenesis, apoptosis, and host defenses. Numerous studies have linked altered MMP expression in different human cancers with poor disease prognosis where up-regulation of MMPs are associated with enhanced cancer cell invasion. ADAMDEC1 is thought to have a central role in dendrite cell functions and their interactions with germinal center T-cells.

The manual screening and GO-analysis results further corroborate that the identified modules are biologically meaningful and that the RCM provides a useful estimate of the covariance.

4.2 Supervised classification

Another application of the RCM is discriminant analysis. As seen below, the estimates obtained can be utilized in supervised learning as a intermediate case between linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA).

Suppose Y is a random variable denoting the classes 1 through k and suppose $\mathbf{x}$ is a random vector of the explanatory variables. Recall that LDA and QDA estimate the class y by maximizing

$$P(Y = y|\mathbf{X} = \mathbf{x}) = \frac{\pi_y f(\mathbf{x}|Y = y)}{\sum_{y'=1}^{k} \pi_{y'} f(\mathbf{x}|Y = y')}$$

over y, where $\mathbf{X}|Y = y$ is assumed to be p-dimensional gaussian distributed, i.e.

$$\mathbf{X}|Y = y \sim \mathcal{N}_p(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y).$$

LDA differs from QDA by the additional assumption that $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_y$ for all classes y. An intermediate classifier of LDA and QDA can thus be constructed by assuming the $\boldsymbol{\Sigma}_y$'s are inverse Wishart distributed as in (1), i.e. $\boldsymbol{\Sigma}_y \sim \mathcal{W}_p^{-1}\big((\nu - p - 1)\boldsymbol{\Sigma}, \nu\big)$, and using the estimates of a common expected $\boldsymbol{\Psi}$ as discussed above. The hierarchical discriminant analysis (HDA) is thus straightforward to implement as

$$\begin{aligned} f(\mathbf{x}|Y = y) &= \int f(\mathbf{x}|\boldsymbol{\Sigma}, Y = y) f(\boldsymbol{\Sigma}|Y = y) d\boldsymbol{\Sigma} \\ &= \frac{\big|\boldsymbol{\Psi}\big|^{\frac{\nu}{2}} \Gamma_p\big(\frac{\nu+1}{2}\big)}{\pi^{-\frac{n}{2}} \big|\boldsymbol{\Psi} + (\mathbf{x} - \boldsymbol{\mu}_y)(\mathbf{x} - \boldsymbol{\mu}_y)^\top\big|^{\frac{\nu+1}{2}} \Gamma_p\big(\frac{\nu}{2}\big)}. \end{aligned}$$

The derivation of which is analogous to Appendix A. The matrix determinant lemma, $|\mathbf{A} + \mathbf{u}\boldsymbol{\Psi}^\top| = (1 + \boldsymbol{\Psi}^\top \mathbf{A}\mathbf{u})|\mathbf{A}|$, can then be applied to simplify the expression and speed up the computations [9]. Note that HDA generalizes the LDA to have a multivariate t distribution. In fact it becomes a multivariate t-distribution when the sample sizes of each class is 1. Multivariate t-distributions have before been considered for discriminant analysis, cf. Andrews and McNicholas [1]. Note that while the classifier is multivariate t-distributed, the estimation procedure assumes correlated observations within each class.

Benchmarking of HDA

We designed four different scenarios to test and identify where HDA can be expected to perform similarly, better, or worse than LDA and QDA as gauged by the misclassification risk. The simulation experiment was inspired by the one seen in Friedman [15].

In all four scenarios, we generated for $p = 5, 10, 20, 35$ a training dataset of $n = 40$ observations belonging to $k = 3$ classes. First, class labels were generated from a multinomial distribution with equal probabilities for each class, $\pi_1 = \pi_2 = \pi_3 = 1/3$. Hence, in each simulation round 13.33 observations were expected in each class. Conditional on the class the observations where drawn i.i.d. from a multivariate gaussian distribution, i.e. $x_i|K = k \sim \mathcal{N}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. The four scenarios consists of different choices of covariance matrices $\boldsymbol{\Sigma}_k$ and mean values $\boldsymbol{\mu}_k$ for each class.

The 3 covariance matrices were chosen to be either (a) equal and spherical, (b) unequal and spherical, (c) equal and highly elliptical, and (d) unequal and highly elliptical. In scenario (a), $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = \mathbf{I}$. In scenario (b), $\boldsymbol{\Sigma}_k = k\mathbf{I}$. In scenario (c), the covariance matrices are equal, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3$, and chosen such that the square root of the d eigenvalues are equidistant on the interval from 10 to 1 and a randomly (uniformly) oriented orthonormal basis is used for all components [15]. In scenario (d), the eigenvalues are chosen as in scenario (c) for all $\boldsymbol{\Sigma}_k$ but the orientation of the orthonormal basis of eigenvectors differs.

In all scenarios expect (c) the mean values were chosen so $\boldsymbol{\mu}_1 = \mathbf{0}, \boldsymbol{\mu}_2 = 3\mathbf{e}_1$, and $\boldsymbol{\mu}_3 = 4\mathbf{e}_2$ where $\mathbf{e}_j$ denotes the j'th basis-vector. In scenario (c), $\boldsymbol{\mu}_1 = \mathbf{0}$ and the remaining mean values are chosen such that the differences project mainly onto the low-variance subspace.

Using the described parameters a training and validation set each of 40 and 100 observations, respectively, were generated. For each method the classifier was trained on the training data followed by classification of the validation data and computation of the misclassification risk. Each simulation setup was repeated 2500 times and the mean and standard error, $\sqrt{\hat{p}(1-\hat{p})/n}$, of the misclassification risks were computed. Table 2 shows the results of the simulation experiments.

In order to be able to evaluate $P(Y = y|\mathbf{X} = \mathbf{x})$ in cases where the sample covariance matrix is not invertible, a small constant was added to the diagonal to allow for stable inversion similar to Friedman [15]. In the given setup, this

Table 2: The estimated misclassification risk for the different scenarios. The minimum and maximum misclassification risks are highlighted in green and red, respectively.

$\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_3$	**Mean misclassification risk (sd)**			
	$p = 5$	$p = 10$	$p = 20$	$p = 35$
Equal, spherical				
LDA	.306 (.009)	.337 (.009)	.401 (.010)	.555 (.010)
QDA	.370 (.010)	.546 (.010)	.666 (.009)	.664 (.009)
HDA	.305 (.009)	.337 (.009)	.401 (.010)	.545 (.010)
Unequal, spherical				
LDA	.378 (.010)	.403 (.010)	.456 (.010)	.609 (.010)
QDA	.418 (.010)	.569 (.010)	.667 (.009)	.666 (.009)
HDA	.376 (.010)	.402 (.010)	.455 (.010)	.569 (.010)
Equal, ellipsoidal				
LDA	.026 (.003)	.008 (.002)	.005 (.001)	.174 (.008)
QDA	.080 (.005)	.432 (.010)	.667 (.009)	.666 (.009)
HDA	.026 (.003)	.008 (.002)	.005 (.001)	.096 (.006)
Unequal, ellipsoidal				
LDA	.563 (.010)	.612 (.010)	.617 (.010)	.655 (.010)
QDA	.209 (.008)	.444 (.010)	.667 (.009)	.668 (.009)
HDA	.550 (.010)	.608 (.010)	.616 (.010)	.655 (.010)

modification was only necessary for QDA in the low-dimensional cases.

In the equal and spherical case (a), HDA yields almost identical results to LDA, which both unsurprisingly outperform QDA. Also perhaps expected, the difference between LDA and QDA is less prominent for low dimensional spaces. The same holds true for the unequal and spherical case (b) and (c).

Most interestingly, HDA is seen to always perform at least as good as LDA in all scenarios. HDA is also consistently superior for the large dimensional tests. The largest gain from HDA to LDA was seen in the high-dimensional scenario (c). This demonstrates HDA as a potentially useful addition to the discriminant analysis toolbox.

5 Concluding remarks

This article provides a basic framework for modeling a common covariance structure across multiple classes or datasets. The straight-forward approaches of using the mean or pooled covariance matrix are seen as moment estimators of the model and the estimate using the EM algorithm is shown to be superior to these simple alternatives. While the improvements are modest, the article demonstrates a potentially advantageous way of modelling the inter-study variability by a hierarchical random effects model. However, the virtue of such a model is not from improvement in accuracy alone. Also desirable is the explicit and interpretable quantification of the inter-study variance. If $\hat{\nu}$ is estimated

to be large, the studies exhibit a largely common covariance structure, and vice-versa when $\hat{\nu}$ is small. We have provided the ICC for the RCM as an attempt to aid in the interpretation of ν. Also provided is the basic framework for testing if study heterogeneity is present. However, the proposed testing is computationally demanding and only feasible when p is sufficiently small. This could e.g. be overcome by an improved and faster fitting procedures or by deriving the distribution of $\hat{\nu}$ under the null hypothesis. Yet the latter is seemingly intractable as $\hat{\nu}$ is a very complex function of the data. The fact that the null-hypothesis lies on the edge of parameter space also seems to constrain the feasibility of deriving such a distribution.

Additionally, one might question whether the added utility of the ν parameter is an obvious relaxation of covariance homogeneity. For example, it is unclear how large a proportion a *single* extra parameter can explain of the inter-study variance. Hence, the present work should be considered a first step in the direction of explicitly modeling the inter-study variation of covariances.

As demonstrated, combining multiple studies can yield a sufficiently large total sample size $n_\bullet$ that allows for estimation of large covariance matrices without the use of regularization. The generalization of the model to $p \gg n_\bullet$ is extremely interesting though out of scope for this article. We believe this work could be further enriched by combining the method with regularized estimation.

The recent advances in such regularized techniques which allows for analysis of large covariance matrices has unfortunately diminished the focus on collecting an adequate number of samples. The technically possible estimates for extreme n/p ratios does not necessarily imply that a good estimate is achieved. For example, while non-zero entries often can be accurately recalled in graphical LASSO, actual estimates of the covariances (or precisions) can still be heavily biased. Large sample-sizes are still needed to achieve unbiased estimates of the covariance due to the bias-variance trade-off. Therefore, an increased focus should also be appointed to efficiently aggregating datasets and achieving sufficiently large sample sizes to allow for stable and unbiased estimation of covariance matrices.

As an addition to the discriminant analysis toolbox, we recognize that further and more sophisticated simulation experiments are needed to explore scenarios where HDA should be considered a serious alternative.

Acknowledgments

We thank Martin Raussen and Jon Johnsen for their assistance on some of the mathematical proofs. The helpful statistical comments from Steffen Falgreen were much appreciated. The technical assistance from Alexander Schmitz, Julie S. Bødker, Ann-Maria Jensen, Louise H. Madsen, and Helle Høholt is also greatly appreciated.

Supplement A

Documents for reproducibility. The documents and other needed files to perform the analyses to reproduce this article. See the `README` file herein. http://people.math.aau.dk/~abilgrau/RCM/SuppA/

Supplement B

Identified modules and GO analysis. Tables of gene module memberships, auxiliary information, and the significant GO-terms for each identified module. http://people.math.aau.dk/~abilgrau/RCM/SuppB/

References

[1] J. L. Andrews and P. D. McNicholas. Model-based Clustering, Classification, and Discriminant Analysis via Mixtures of Multivariate t-distributions. *Statistics and Computing*, 22(5):1021–1029, Aug. 2011. ISSN 0960-3174. doi: 10.1007/s11222-011-9272-x.

[2] A. J. Butte, P. Tamayo, D. Slonim, T. R. Golub, and I. S. Kohane. Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. *PNAS*, 97(22):12182–6, Oct. 2000. ISSN 0027-8424. doi: 10.1073/pnas.220392197.

[3] J. K. Choi, U. Yu, S. Kim, and O. J. Yoo. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 19(Suppl 1): i84–i90, July 2003. ISSN 1367-4803. doi: 10.1093/bioinformatics/btg1010.

[4] R. D. Cook and L. Forzani. On the mean and variance of the generalized inverse of a singular wishart matrix. *Electronic Journal of Statistics*, 5: 146–158, 2011. ISSN 1935-7524. doi: 10.1214/11-EJS602.

[5] M. Dai, P. Wang, A. D. Boyd, G. Kostov, B. Athey, E. G. Jones, W. E. Bunney, R. M. Myers, T. P. Speed, H. Akil, S. J. Watson, and F. Meng. Evolving gene/transcript definitions significantly alter the interpretation of genechip data. *Nucleic Acids Research*, 33(20):e175, Jan. 2005. ISSN 1362-4962. doi: 10.1093/nar/gni179.

[6] A. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972.

[7] A. Dempster, N. Laird, and D. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 39(1):1–38, 1977.

[8] R. DerSimonian and N. Laird. Meta-analysis in Clinical Trials. *Controlled Clinical Trials*, 7(3):177–88, Sept. 1986. ISSN 0197-2456.

[9] J. Ding and A. Zhou. Eigenvalues of rank-one updated matrices with some applications. *Applied Mathematics Letters*, 20(12):1223–1226, Dec. 2007. ISSN 08939659. doi: 10.1016/j.aml.2006.11.016.

[10] K. Dybkær, M. Bøgsted, S. Falgreen, J. S. Bødker, M. K. Kjeldsen, A. Schmitz, A. E. Bilgrau, Z. Y. Xu-Monette, L. Li, K. S. Bergkvist, M. B. Laursen, M. Rodrigo-Domingo, S. C. Marques, S. B. Rasmussen, M. Nyegaard, M. Gaihede, M. B. Møller, R. J. Samworth, R. D. Shah, P. Johansen, T. C. El-Galaly, K. H. Young, and H. E. Johnsen. A diffuse large B-cell lymphoma classification system that associates normal B-cell subset phenotypes with prognosis. *Journal Of Clinical Oncology*, 33(12): 1379–1388, 2015.

[11] D. Eddelbuettel and R. François. **Rcpp**: Seamless R and C++ Integration. *Journal of Statistical Software*, 40(8), 2011.

[12] R. François, D. Eddelbuettel, and D. Bates. **RcppArmadillo***:* **Rcpp** *Integration for Armadillo Templated Linear Algebra Library*, 2012. URL http://CRAN.R-project.org/package=RcppArmadillo. R package version 0.3.6.1.

[13] R. A. Francescone, S. Scully, M. Faibish, S. L. Taylor, D. Oh, L. Moral, W. Yan, B. Bentley, and R. Shao. Role of YKL-40 in the angiogenesis, radioresistance, and progression of glioblastoma. *Journal of Biological Chemistry*, 286(17):15332–15343, 2011.

[14] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–41, July 2008. ISSN 1468-4357. doi: 10.1093/biostatistics/kxm045.

[15] J. H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175, 1989.

[16] T. M. Fruchterman and E. M. Reingold. Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11):1129–1164, 1991.

[17] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry. affy—analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3):307–315, 2004. ISSN 1367-4803. doi: http://dx.doi.org/10.1093/bioinformatics/btg405.

[18] D. Holten. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *Visualization and Computer Graphics, IEEE Transactions on*, 12(5):741–748, 2006.

[19] S. Horvath. *Weighted Network Analysis: Applications in Genomics and Systems Biology*. Springer, 2011.

[20] International Lymphoma Study Group. A clinical evaluation of the international lymphoma study group classification of non-hodgkin's lymphoma. *Blood*, 89(11):3909–3918, June 1997. The Non-Hodgkin's Lymphoma Classification Project.

[21] H. K. Khalil. *Nonlinear Systems*. Prentice Hall, 2002. ISBN 0-13-067389-7.

[22] F.-H. Lin, S.-Y. Tsai, R. Otazo, A. Caprihan, L. L. Wald, J. W. Belliveau, and S. Posse. Sensitivity-Encoded (SENSE) Proton Echo-Planar Spectroscopic Imaging (PEPSI) in the Human Brain. *Magnetic Resonance in Medicine*, 57(2):249–57, Feb. 2007. ISSN 0740-3194. doi: 10.1002/mrm.21119.

[23] K. Petersen and M. Pedersen. *The Matrix Cookbook*. Technical University of Denmark, Technical Manual, 2008. URL http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=3274.

[24] B. Phipson and G. K. Smyth. Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(39), 2010.

[25] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL http://www.R-project.org/. ISBN 3-900051-07-0.

[26] P. E. Shrout and J. L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420, 1979.

[27] W. N. van Wieringen and C. F. W. Peeters. Ridge estimation of inverse covariance matrices from high-dimensional data. Submitted to *Computational Statistics & Data Analysis, arXiv:1403.0904v3*, 2015.

[28] D. von Rosen. Moments for the Inverted Wishart Distribution. *Scandinavian Journal of Statistics*, 15(2):97–109, 1988.

[29] Y. Xie. *Dynamic Documents with* R *and* **knitr**. CRC Press, 2013. ISBN 9781482203530.

A Marginalization of the covariance

This section shows the marginalization over $\mathbf{\Sigma}$ in (4). For ease of notation we drop the subscript i on $\mathbf{\Sigma}_i$, $\mathbf{X}_i$, $\mathbf{S}_i = \mathbf{X}_i\mathbf{X}_i^\top$, and n_i. By the model assumptions,

$$
\begin{aligned}
f(\mathbf{X}|\mathbf{\Psi},\nu) &= \int f(\mathbf{X}|\mathbf{\Sigma}) f(\mathbf{\Sigma}|\mathbf{\Psi},\nu) d\mathbf{\Sigma} \\
&= \int \left[\prod_{j=1}^{n} (2\pi)^{-\frac{p}{2}} |\mathbf{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}\operatorname{tr}(\mathbf{x}_{ij}\mathbf{x}_{ij}^\top \mathbf{\Sigma}^{-1})}\right] \frac{\left|\mathbf{\Psi}\right|^{\frac{\nu}{2}}}{2^{\frac{\nu p}{2}}\Gamma_p\left(\frac{\nu}{2}\right)} |\mathbf{\Sigma}|^{-\frac{\nu+p+1}{2}} e^{-\frac{1}{2}\operatorname{tr}\left(\mathbf{\Psi}\mathbf{\Sigma}^{-1}\right)} \, d\mathbf{\Sigma} \\
&= (2\pi)^{-\frac{np}{2}} \frac{\left|\mathbf{\Psi}\right|^{\frac{\nu}{2}}}{2^{\frac{\nu p}{2}}\Gamma_p\left(\frac{\nu}{2}\right)} \int |\mathbf{\Sigma}|^{-\frac{n}{2}} e^{-\frac{1}{2}\operatorname{tr}(\mathbf{S}\mathbf{\Sigma}^{-1})} |\mathbf{\Sigma}|^{-\frac{\nu+p+1}{2}} e^{-\frac{1}{2}\operatorname{tr}\left(\mathbf{\Psi}\mathbf{\Sigma}^{-1}\right)} \, d\mathbf{\Sigma} \\
&= \frac{\left|\mathbf{\Psi}\right|^{\frac{\nu}{2}}}{\pi^{\frac{np}{2}} 2^{\frac{(\nu+n)p}{2}}\Gamma_p\left(\frac{\nu}{2}\right)} \int |\mathbf{\Sigma}|^{-\frac{(\nu+n)+p+1}{2}} e^{-\frac{1}{2}\operatorname{tr}\left(\left(\mathbf{\Psi}+\mathbf{S}\right)\mathbf{\Sigma}^{-1}\right)} \, d\mathbf{\Sigma}.
\end{aligned}
$$

The integrand can be recognized as a unnormalized inverse Wishart pdf of the distribution $\mathcal{W}^{-1}\big(\mathbf{\Psi}+\mathbf{S}, \nu+n\big)$, and so the integral evaluates to the reciprocal value of the normalizing constant in that density. Thus,

$$
f(\mathbf{X}|\mathbf{\Psi},\nu) = \frac{\left|\mathbf{\Psi}\right|^{\frac{\nu}{2}}}{\pi^{\frac{np}{2}} 2^{\frac{(\nu+n)p}{2}}\Gamma_p\left(\frac{\nu}{2}\right)} \frac{2^{\frac{(\nu+n)p}{2}}\Gamma_p\left(\frac{\nu+n}{2}\right)}{\left|\mathbf{\Psi}+\mathbf{S}\right|^{\frac{\nu+n}{2}}} = \frac{\left|\mathbf{\Psi}\right|^{\frac{\nu}{2}}\Gamma_p\left(\frac{\nu+n}{2}\right)}{\pi^{\frac{np}{2}}\left|\mathbf{\Psi}+\mathbf{S}\right|^{\frac{\nu+n}{2}}\Gamma_p\left(\frac{\nu}{2}\right)}. \tag{13}
$$

Using the matrix determinant lemma and $\mathbf{S} = \mathbf{X}^\top\mathbf{X}$, this can be further simplified to

$$
f(\mathbf{X}|\mathbf{\Psi},\nu) = \frac{\Gamma_p\left(\frac{\nu+n}{2}\right)}{\pi^{\frac{np}{2}}\left|\mathbf{I}+\mathbf{X}\mathbf{\Psi}^{-1}\mathbf{X}^\top\right|^{\frac{\nu+n}{2}}\left|\mathbf{\Psi}\right|^{\frac{n}{2}}\Gamma_p\left(\frac{\nu}{2}\right)},
$$

which can help to speed-up computations.

B Proofs

B.1 Non-concavity of the log-likelihood

The likelihood function is not log-concave in general. This section analyses the (non)-concavity of the log-likelihood function given in (4). More precisely, the following two propositions are proved.

Proposition 1 (Non-concavity in $\mathbf{\Psi}$)
For fixed ν, the log-likelihood function (4) is not concave in $\mathbf{\Psi}$.

Proposition 2 (Concavity in ν)
For fixed positive definite $\mathbf{\Psi}$, the log-likelihood function (4) is concave in ν.

Proof of Proposition 1. Assume ν is fixed and consider only the terms involving $\mathbf{\Psi}$ in (4). We reduce to the one-dimensional case where

$$\ell(\psi) = \frac{k\nu}{2}\log(\psi) - \sum_{i=1}^{k}\frac{\nu+n_i}{2}\log(\psi + x_i^2),$$

which implies

$$\ell'(\psi) = \frac{k\nu}{2}\frac{1}{\psi} - \sum_{i=1}^{k}\frac{\nu+n_i}{2}\frac{1}{\psi+x_i^2} \text{ and } \ell''(\psi) = -\frac{k\nu}{2}\frac{1}{\psi^2} + \sum_{i=1}^{k}\frac{\nu+n_i}{2}\frac{1}{(\psi+x_i^2)^2}.$$

It is straightforward to show there exists a value for ψ, n_i and ν for which $\ell''(\psi) > 0$. Since the second derivative is not always negative the log-likelihood ℓ is not log-concave. □

Proof of Proposition 2. Consider the terms involving ν. Clearly, the mixed terms involving both ν and $\mathbf{\Psi}$ are log-linear in ν and hence log-concave. We thus restrict our attention to the remaining terms not dependent on $\mathbf{\Psi}$. The sum of these terms are concave in ν, since

$$\log\Gamma_p\left(\frac{\nu+n_i}{2}\right) - \log\Gamma_p\left(\frac{\nu}{2}\right) = \log\frac{\Gamma_p\left(\frac{\nu+n_i}{2}\right)}{\Gamma_p\left(\frac{\nu}{2}\right)} = \sum_{j=1}^{p}\log\frac{\Gamma\left(\frac{\nu+1-j}{2}+\frac{n_i}{2}\right)}{\Gamma\left(\frac{\nu+1-j}{2}\right)}.$$

which can be seen to be concave since $n_i \geq 1$ for all i and $h(x) = \log\left(\frac{\Gamma(x+a)}{\Gamma(x)}\right)$ is concave for all $x > 0$ and $a > 0$. The concavity of h is easily seen by the fact that $h''(x) = \psi(x+a) - \psi(x) < 0$, where $\psi(\cdot)$ is the tri-gamma function. The tri-gamma function is a well-known monotonically decreasing function. Hence, the log-likelihood is log-concave in ν. □

log-convexity of the multivariate gamma function

The multivariate gamma function Γ_p is log-convex as can be seen using the following characterization,

$$\Gamma_p(t) = \pi^{\frac{1}{2}\binom{p}{2}}\prod_{j=1}^{p}\Gamma\left(t+\frac{1-j}{2}\right) \text{ where } \Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx. \tag{14}$$

From this

$$\log\Gamma_p(t) = \frac{1}{2}\binom{p}{2}\log\pi + \sum_{j=1}^{p}\Gamma\left(t+\frac{1-j}{2}\right), \tag{15}$$

which is convex since Γ is log-convex and a sum of convex functions is convex. Hence Γ_p is also log-convex.

B.2 Existence and uniqueness of likelihood maxima

This section proves Lemmas 1 and 2 which imply Proposition 3.

Before we state the lemmas, the proposition, and their proofs, we see that the reparameterisation of the RCM is irrelevant. Consider the log-likelihood in (4) assuming ν fixed. The log-likelihood obey

$$2\ell(\mathbf{\Psi}) = c + k\nu \log\left|\mathbf{\Psi}\right| - \sum_{a=1}^{k}(n_a+\nu)\log\left|\mathbf{\Psi}+\mathbf{S}_a\right|. \tag{16}$$

Notice, that this equation also holds in the reparameterization. Here we have

$$\begin{aligned} 2\ell(\mathbf{\Sigma}) &= c + k\nu \log\left|(\nu-p-1)\mathbf{\Sigma}\right| - \sum_{a=1}^{k}(n_a+\nu)\log\left|(\nu-p-1)\mathbf{\Sigma}+\mathbf{S}_a\right| \\ &= c' + k\nu\log\left|\mathbf{\Sigma}\right| - \sum_{a=1}^{k}(n_a+\nu)\log\left|\mathbf{\Sigma}+(\nu-p-1)^{-1}\mathbf{S}_a\right|. \end{aligned}$$

Since $(\nu-p-1)^{-1}\mathbf{S}_a$ is only dependent on data (when ν is fixed) we can set $(\nu-p-1)^{-1}\mathbf{S}_a := \mathbf{S}_a$. Without loss of generality we can therefore consider (16) in the following.

Proposition 3 (Existence and uniqueness)
The log-likelihood (4) has a unique maximum in $\mathbf{\Psi}$ for fixed ν and $n_\bullet = \sum_{a=1}^{k} n_a \geq p$.

Proof of Proposition 3. We first prove the existence of the maximum. By Lemma 1 and the continuity of ℓ, the set $\left\{\mathbf{\Psi}\middle|\ell(\mathbf{\Psi}) \geq \ell(\mathbf{\Psi}^*)\right\}$ is bounded and closed and thus compact for any $\mathbf{\Psi}^* \succ 0$. The existence of a maximum follows from the extreme value theorem by the continuity of ℓ. A stationary point exists due to Rolle's theorem and the differentiability of ℓ.

Next, we show the uniqueness of the maximum. Suppose there exists countably many stationary points $\mathbf{\Psi}_1, \mathbf{\Psi}_2, \ldots$. By Lemma 1, $\ell(\mathbf{\Psi})$ has a finite upper bound given by the value of the log-likelihood in those points. All gradient curves (that is, solution curves to $\dot{\mathbf{\Psi}}(t) = \nabla\ell(\mathbf{\Psi}(t))$) must then converge toward exactly one of the stationary points where ℓ monotonically increases along each curve. Define the basin of attraction

$$A_i = \left\{\mathbf{\Psi}_0 \in \mathcal{S}_+ \middle| \mathbf{\Psi}(0) = \mathbf{\Psi}_0,\ \lim_{t\to\infty}\mathbf{\Psi}(t) = \mathbf{\Psi}_i\right\},$$

associated to each stationary point $\mathbf{\Psi}_i$. The basin of attraction is open if $\mathbf{\Psi}_i$ is a maximum [21, Lemma 4.1]. By Lemma 2, $\mathbf{\Psi}_i$ is a always a maximum and hence all A_i are open sets in the set of all positive semi-definite matrices $\mathcal{S}_+$. This partitions the space $\mathcal{S}_+$ into N disjoint, non-empty, open sets. However, this is only possible if $A_i = A_j = \mathcal{S}_+$ for all i and j and thus there is only a single basin of attraction and maximum of ℓ. □

Lemma 1
If there exists an eigenvalue λ_t of $\boldsymbol{\Psi}_t$ such that $\lambda_t \to 0$ or $\lambda_t \to \infty$, then $\ell(\boldsymbol{\Psi}_t) \to -\infty$ for ν fixed and $n_\bullet = \sum_{a=1}^{k} n_a \geq p$.

Proof of Lemma 1. Assume the hypothesis of the lemma and consider the expression given in (16). If $\lambda_t \to \infty$ then

$$\begin{aligned}\ell(\boldsymbol{\Psi}_t) &= \frac{k\nu}{2}\log\left|\boldsymbol{\Psi}_t\right| - \sum_{i=1}^{k}\frac{\nu+n_i}{2}\log|\boldsymbol{\Psi}_t + \mathbf{S}_i| \\ &\leq \frac{k\nu}{2}\log|\boldsymbol{\Psi}_t| - \sum_{i=1}^{k}\frac{\nu+n_i}{2}\log|\boldsymbol{\Psi}_t| = -\frac{n_\bullet}{2}\log|\boldsymbol{\Psi}_t| \to -\infty.\end{aligned}$$

This proves the first case where the largest eigenvalue diverge to infinity. Suppose $\lambda_t \to 0$ and let $C = \sum_{i=1}^{k}\frac{\nu+n_i}{2} = \frac{k\nu}{2} + \frac{n_\bullet}{2}$, then (16) can be expressed as

$$\ell(\boldsymbol{\Psi}_t) = \frac{k\nu}{2}\log|\boldsymbol{\Psi}_t| - C\sum_{i=1}^{k}\frac{\nu+n_i}{2C}\log|\boldsymbol{\Psi}_t + \mathbf{S}_i|.$$

Since $\log|\cdot|$ is concave and the above sum is a convex combination, we have

$$\ell(\boldsymbol{\Psi}_t) \leq \frac{k\nu}{2}\log|\boldsymbol{\Psi}_t| - C\log\left|\boldsymbol{\Psi}_t + \sum_{i=1}^{k}\frac{\nu+n_i}{2C}\mathbf{S}_i\right|.$$

Clearly, the first term goes to $-\infty$ whenever an eigenvalue $\lambda_t \to 0$. The matrix in the second term is almost surely positive definite when $n_\bullet = \sum_{a=1}^{k} x_a \geq p$ and the log determinant will converge to some constant. Hence, if $\lambda_t \to 0$ then

$$\ell(\boldsymbol{\Psi}_t) \leq \frac{k\nu}{2}\log|\boldsymbol{\Psi}_t| + C_2 \to -\infty,$$

which completes the proof. □

Lemma 2
If $n_\bullet \geq p$ and ν is fixed then the Hessian of the log-likelihood (4) is negative definite in all stationary points.

Proof of Lemma 2. We show the conclusion of the Lemma directly by differentiation of ℓ w.r.t. $\boldsymbol{\Psi}$. To do so, the matrix cookbook by Petersen and Pedersen [23] is a useful reference. In particular, see equations (41, p. 8) and

(59, p. 9) and pages 14 and 52-53. We first compute expressions for the first and second order derivatives.

First order derivatives. From the log-likelihood expression, we compute the first order derivative $\nabla_{\mathbf{\Psi}} 2\ell(\mathbf{\Psi})$ which is the matrix-valued function where each entry is given by

$$\frac{\partial 2\ell}{\partial \Psi_{ij}} = k\nu \operatorname{tr}\left(\mathbf{E}^{ij}\mathbf{\Psi}^{-1}\right) - \sum_{a=1}^{k}(\nu + n_a)\operatorname{tr}\left(\mathbf{E}^{ij}\left(\mathbf{\Psi} + \mathbf{S}_a\right)^{-1}\right). \tag{17}$$

and $\mathbf{E}^{ij}$ is a matrix with ones at entries (i,j) and (j,i) and zeros elsewhere. This $\mathbf{E}^{ij}$ is introduced as the derivative is not straight-forward because of the symmetric structure of $\mathbf{\Psi}$. Had $\mathbf{\Psi}$ been unstructured, then $\frac{\partial}{\partial \mathbf{\Psi}} \log|\mathbf{\Psi}| = \mathbf{\Psi}^{-1}$. However, when $\mathbf{\Psi}$ is symmetric we have that $\frac{\partial}{\partial \Psi_{ij}} \log|\mathbf{\Psi}| = \operatorname{tr}(\mathbf{E}^{ij}\mathbf{\Psi}^{-1})$ which is the same as $\frac{\partial}{\partial \mathbf{\Psi}} \log|\mathbf{\Psi}| = 2\mathbf{\Psi}^{-1} - \mathbf{\Psi}^{-1} \circ \mathbf{I}$ where $\circ$ denotes the Hadamard product [23, eq. (43) and (141)].

The first order derivative lives in a $\binom{p+1}{2}$-dimensional vector space with basis vectors $\mathbf{E}^{ij}$ indexed by (i,j), $i \leq j$.

Second order derivatives. We proceed with the second order derivative $\nabla^2_{\mathbf{\Psi}} 2\ell(\mathbf{\Psi})$ with entries given by

$$\begin{aligned}\frac{\partial^2 2\ell}{\partial \Psi_{kl}\partial \Psi_{ij}} &= -k\nu \operatorname{tr}\left(\mathbf{E}^{ij}\mathbf{\Psi}^{-1}\mathbf{E}^{kl}\mathbf{\Psi}^{-1}\right) \\ &\quad + \sum_{a=1}^{k}(\nu + n_a)\operatorname{tr}\left(\mathbf{E}^{ij}\left(\mathbf{\Psi} + \mathbf{S}_a\right)^{-1}\mathbf{E}^{kl}\left(\mathbf{\Psi} + \mathbf{S}_a\right)^{-1}\right),\end{aligned}$$

obtained by differentiation of (17) using $\frac{\partial}{\partial \Psi_{ij}}\mathbf{\Psi}^{-1} = -\mathbf{\Psi}^{-1}\mathbf{E}^{ij}\mathbf{\Psi}^{-1}$ [23, eq. (40)] and the linearity of the trace operator.

The second order derivative is a $\binom{p+1}{2} \times \binom{p+1}{2}$-dimensional matrix indexed by (i,j) and (k,l), $i \leq j$, $k \leq l$.

Negative definiteness of stationary points. With the above expressions we now show that the Hessian matrix is negative definite in all stationary points. Let $\mathbf{Y} = \sum_{(i,j)} y_{ij}\mathbf{E}^{ij}$ be an arbitrary symmetric matrix in the vector space where $\mathbf{Y} \neq \mathbf{0}$. In our vector space we need to show that

$$\sum_{i\leq j, k\leq j} Y_{ij}\left(\nabla^2_{\mathbf{\Psi}} 2\ell(\mathbf{\Psi})\right)_{(i,j),(k,l)} Y_{kl} < 0$$

holds in every stationary point analogous to $\mathbf{z}^\top\mathbf{A}\mathbf{z} = \sum_{ij} A_{ij}z_i z_j < 0$. From the second derivative, this amounts to showing that in every stationary point,

$$-k\nu \operatorname{tr}\left(\mathbf{Y}\mathbf{\Psi}^{-1}\mathbf{Y}\mathbf{\Psi}^{-1}\right) + \sum_{a=1}^{k}(\nu + n_a)\operatorname{tr}\left(\mathbf{Y}\left(\mathbf{\Psi} + \mathbf{S}_a\right)^{-1}\mathbf{Y}\left(\mathbf{\Psi} + \mathbf{S}_a\right)^{-1}\right) < 0. \tag{18}$$

Now, by the positive-definiteness of $\mathbf{\Psi}$, let

$$\mathbf{Y} := \mathbf{\Psi}^{-\frac{1}{2}}\mathbf{Y}\mathbf{\Psi}^{-\frac{1}{2}} \text{ and}$$
$$\mathbf{S}_a := \mathbf{\Psi}^{-\frac{1}{2}}\mathbf{S}_a\mathbf{\Psi}^{-\frac{1}{2}},$$

and thus without loss of generality we can assume that $\mathbf{\Psi} = \mathbf{I}$. Hence, the derivative of the likelihood (17) equated to zero, becomes

$$k\nu\mathbf{I} = \sum_a (n_a + \nu)(\mathbf{I} + \mathbf{S}_a)^{-1}$$

which implies (by multiplication by $\mathbf{Y}$ on each side) that every stationary point obey

$$k\nu \operatorname{tr}(\mathbf{Y}^2) = \sum_a (n_a + \nu) \operatorname{tr}\Big(\mathbf{Y}(\mathbf{I} + \mathbf{S}_a)^{-1}\mathbf{Y}\Big). \tag{19}$$

We substitute (19) into (18) to get

$$\sum_a (n_a + \nu) \operatorname{tr}\Big(\mathbf{Y}(\mathbf{I} + \mathbf{S}_a)^{-1}\mathbf{Y}(\mathbf{I} + \mathbf{S}_a)^{-1} - \mathbf{Y}(\mathbf{I} + \mathbf{S}_a)^{-1}\mathbf{Y}\Big)$$
$$= \sum_a (n_a + \nu) \operatorname{tr}\Big(\mathbf{Y}(\mathbf{I} + \mathbf{S}_a)^{-1}\mathbf{Y}\big[(\mathbf{I} + \mathbf{S}_a)^{-1} - \mathbf{I}\big]\Big) < 0.$$

We note that $\mathbf{S}_a = \mathbf{X}_a\mathbf{X}_a^\top$ and

$$(\mathbf{I} + \mathbf{S}_a)^{-1} - \mathbf{I} = -\mathbf{X}_a\big(\mathbf{I} + \mathbf{X}_a^\top\mathbf{X}_a\big)^{-1}\mathbf{X}_a^\top,$$

by the matrix inversion lemma whereby we need to show that

$$\sum_a (n_a + \nu) \operatorname{tr}\Big(\mathbf{Y}(\mathbf{I} + \mathbf{X}_a\mathbf{X}_a^\top)^{-1}\mathbf{Y}\mathbf{X}_a\big(\mathbf{I} + \mathbf{X}_a^\top\mathbf{X}_a\big)^{-1}\mathbf{X}_a^\top\Big) > 0.$$

Since $(\mathbf{I} + \mathbf{X}_a\mathbf{X}_a^\top)^{-1} \succ 0$ we obtain that

$$\mathbf{Y}\mathbf{X}_a(\mathbf{I} + \mathbf{X}_a\mathbf{X}_a^\top)^{-1}\mathbf{X}_a^\top\mathbf{Y} = 0 \quad \text{for } a = 1, ..., k.$$

Again by $(\mathbf{I} + \mathbf{X}_a\mathbf{X}_a^\top)^{-1} \succ 0$ we conclude that $\mathbf{Y}\mathbf{X}_a = 0$ for all $a = 1, ..., k$, i.e. $\mathbf{Y}(\mathbf{X}_1, ..., \mathbf{X}_k) = 0$. If $n_\bullet \geq p$ then almost surely $(\mathbf{X}_1, ..., \mathbf{X}_k)$ has rank p whereby $\mathbf{Y} = 0$. □

C Likelihood of the precision matrix

Suppose we have k i.i.d. realizations, $\mathbf{\Delta}_1, ..., \mathbf{\Delta}_k$, from the Wishart distribution given in model (6). The corresponding log-likelihood can be computed straight-

forwardly:

$$\begin{aligned}\ell(\boldsymbol{\Theta}|\boldsymbol{\Delta}_1,...,\boldsymbol{\Delta}_k) &= \sum_{i=1}^{k} \log \frac{\left|\boldsymbol{\Theta}\right|^{-\frac{\nu}{2}}}{2^{-\frac{\nu p}{2}}\Gamma_p\left(\frac{\nu}{2}\right)} |\boldsymbol{\Delta}_i|^{\frac{\nu-p-1}{2}} e^{-\frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Theta}^{-1}\boldsymbol{\Delta}_i\right)} \\ &= c + \sum_{i=1}^{k}\left(-\frac{\nu}{2}\log\left|\boldsymbol{\Theta}\right| - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Theta}^{-1}\boldsymbol{\Delta}_i\right)\right) \\ &= c - \frac{\nu k}{2}\left(\log|\boldsymbol{\Theta}| + \operatorname{tr}\left(\boldsymbol{\Theta}^{-1}\frac{1}{\nu k}\sum_{i=1}^{k}\boldsymbol{\Delta}_i\right)\right).\end{aligned}$$

The last expression is to be maximized with respect to $\boldsymbol{\Theta}$ and can be recognized as the MLE problem in a multivariate Gaussian distribution. Hence, $\boldsymbol{\Theta} = \frac{1}{k\nu}\sum_{i=1}^{k}\boldsymbol{\Delta}_i$, is the MLE in this model.

D Approximate MLE

To find the maximizing parameters we differentiate (4) w.r.t. $\boldsymbol{\Psi}$ and equate to zero while assuming ν known and constant. The first order derivative can be seen in equation (17). Equating to zero yields

$$\begin{aligned}\mathbf{0} &= \frac{k\nu}{2}\boldsymbol{\Psi}^{-1} - \sum_{i=1}^{k}\frac{\nu+n_i}{2}(\boldsymbol{\Psi}+\mathbf{S}_i')^{-1} \\ &= \frac{k\nu}{2}\boldsymbol{\Psi}^{-1} - \sum_{i=1}^{k}\frac{\nu+n_i}{2}\left(\mathbf{I}+\boldsymbol{\Psi}^{-1}\mathbf{S}_i\right)^{-1}\boldsymbol{\Psi}^{-1}.\end{aligned} \tag{20}$$

This implies $k\nu\mathbf{I} - \sum_{i=1}^{k}(\nu+n_i)\left(\mathbf{I}-(-\boldsymbol{\Psi}^{-1}\mathbf{S}_i)\right)^{-1} = \mathbf{0}$ which can be rewritten as

$$k\nu\mathbf{I} - \sum_{i=1}^{k}(\nu+n_i)\sum_{l=0}^{\infty}\left(-\boldsymbol{\Psi}^{-1}\mathbf{S}_i\right)^{l} = \mathbf{0},$$

by the Neumann series $\left((\mathbf{I}+\mathbf{A})^{-1} = \sum_{l=0}^{\infty}\mathbf{A}^l\right)$ provided that $\lim_{l\to\infty}(\mathbf{I} - \boldsymbol{\Psi}^{-1}\mathbf{S}_i)^l = \mathbf{0}$ for all i. This holds if the eigenvalues of $\boldsymbol{\Psi}^{-1}\mathbf{S}_i$ are less than 1. We approximate by the first order expansion ($l=1$), and

$$\mathbf{0} = k\nu\mathbf{I} - \sum_{i=1}^{k}(\nu+n_i)(\mathbf{I}-\boldsymbol{\Psi}^{-1}\mathbf{S}_i) = -n_\bullet\mathbf{I} + \boldsymbol{\Psi}^{-1}\sum_{i=1}^{k}(\nu+n_i)\mathbf{S}_i$$

where $n_\bullet = \sum_{i=1}^{k} n_i$ is the total number of observations. This implies

$$\boldsymbol{\Psi}^{-1}\sum_{i=1}^{k}(\nu+n_i)\mathbf{S}_i = n_\bullet\mathbf{I}$$

which suggests the estimators

$$\hat{\boldsymbol{\Psi}}_{\text{MLE}} = \frac{\sum_{i=1}^{k}(\nu + n_i)\mathbf{S}_i}{n_\bullet} \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}_{\text{MLE}} = \frac{\sum_{i=1}^{k}(\nu + n_i)\mathbf{S}_i}{(\nu - p - 1)n_\bullet}. \tag{21}$$

These estimates are seen to correspond to a weighted sum of the scatter matrices.

Paper IV

Targeted Fused Ridge Estimation of Inverse Covariance Matrices from Multiple High-Dimensional Data Classes

Anders Ellern Bilgrau*, Carel F.W. Peeters*, Poul Svante Eriksen,
Martin Bøgsted, and Wessel N. van Wieringen
*Shared first authorship

Submitted to
Journal of Machine Learning Research, 2015.

Preface: This work attacks the general problem of Paper III from a more contemporary point of view and allows for high-dimensional data. It too seeks to estimate the (inverse) covariance matrices of multiple groups, however with a regularized approach. The method borrows information across groups by incorporating a *fusion* term in the penalty which shrinks the group-estimates toward each other—a method dubbed *fusion* as the estimates are *fused* into one for large penalties. The paper thus explore the usage of a *fused* version of the ridge (ℓ_2) penalty which enjoys some alternative desirable properties than the highly popular graphical lasso (ℓ_1) penalty and its fused version.

The work was prompted by Carel F.W. Peeters and Wessel N. van Wieringen as a targeted version of Price et al. [33] in relation to their previous work [41]. However, the fused ridge penalty was quickly generalized into its novel and very flexible present form.

A large part of the effort went into improving the robustness and speed of the estimation procedure of the non-fused ridge estimator of van Wieringen and Peeters [41]. This was necessary as fast estimators are needed for the fused counterpart and indeed even more so when the optimal penalty parameters need to be determined. This included rewriting much of the code base for **rags2ridges** in C++ for which I earned coauthership of the package from v2.0 and forward. The **rags2ridges** is available at

`http://cran.r-project.org/package=rags2ridges`
`https://github.com/CFWP/rags2ridges`

This paper also makes use of the **DLBCLdata**-package (Package II).

 The layout has been revised.

Targeted Fused Ridge Estimation of Inverse Covariance Matrices from Multiple High-Dimensional Data Classes

Abstract

We consider the problem of jointly estimating multiple inverse covariance matrices from high-dimensional data consisting of distinct classes. An ℓ_2-penalized maximum likelihood approach is employed. The suggested approach is flexible and generic, incorporating several other ℓ_2-penalized estimators as special cases. In addition, the approach allows specification of target matrices through which prior knowledge may be incorporated and which can stabilize the estimation procedure in high-dimensional settings. The result is a targeted fused ridge estimator that is of use when the precision matrices of the constituent classes are believed to chiefly share the same structure while potentially differing in a number of locations of interest. It has many applications in (multi)factorial study designs. We focus on the graphical interpretation of precision matrices with the proposed estimator then serving as a basis for integrative or meta-analytic Gaussian graphical modeling. Situations are considered in which the classes are defined by datasets and subtypes of diseases. The performance of the proposed estimator in the graphical modeling setting is assessed through extensive simulation experiments. Its practical usability is illustrated by the differential network modeling of 12 large-scale gene expression datasets of diffuse large B-cell lymphoma subtypes. The estimator and its related procedures are incorporated into the R-package **rags2ridges**.

1 Introduction

High-dimensional data are ubiquitous in modern statistics. Consequently, the fundamental problem of estimating the covariance matrix or its inverse (the precision matrix) has received renewed attention. Suppose we have n i.i.d. observations of a p-dimensional variate distributed as $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The Gaussian log-likelihood parameterized in terms of the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is then given by:

$$\mathcal{L}(\boldsymbol{\Omega}; \mathbf{S}) \propto \ln|\boldsymbol{\Omega}| - \text{tr}(\mathbf{S}\boldsymbol{\Omega}), \tag{1}$$

where $\mathbf{S}$ is the sample covariance matrix. When $n > p$ the maximum of (1) is attained at the maximum likelihood estimate (MLE) $\hat{\boldsymbol{\Omega}}^{\text{ML}} = \mathbf{S}^{-1}$. However, in the high-dimensional case, i.e., when $p > n$, the sample covariance matrix $\mathbf{S}$ is singular and its inverse ceases to exist. Furthermore, when $p \approx n$, the sample covariance matrix may be ill-conditioned and the inversion becomes numerically unstable. Hence, these situations necessitate usage of regularization techniques.

Here, we study the simultaneous estimation of numerous precision matrices when multiple classes of high-dimensional data are present. Suppose $\mathbf{y}_{ig}$ is a realization of a p-dimensional Gaussian random vector for $i = 1, \ldots, n_g$ indepen-

dent observations nested within $g = 1, \ldots, G$ classes, each with class-dependent covariance $\boldsymbol{\Sigma}_g$, i.e., $\mathbf{y}_{ig} \sim \mathcal{N}_p(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ for each designated class g. Hence, for each class a dataset consisting of the $n_g \times p$ matrix $\mathbf{Y}_g = [\mathbf{y}_{1g}, \ldots, \mathbf{y}_{n_g g}]^\top$ is observed. Without loss of generality $\boldsymbol{\mu}_g = \mathbf{0}$ can be assumed as each dataset $\mathbf{Y}_g$ can be centered around its column means. The class-specific sample covariance matrix given by

$$\mathbf{S}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} \mathbf{y}_{ig} \mathbf{y}_{ig}^\top = \frac{1}{n_g} \mathbf{Y}_g^\top \mathbf{Y}_g,$$

then constitutes the well-known MLE of $\boldsymbol{\Sigma}_g$ as discussed above. The closely related *pooled* sample covariance matrix

$$\mathbf{S}_\bullet = \frac{1}{n_\bullet} \sum_{g=1}^{G} \sum_{i=1}^{n_g} \mathbf{y}_{ig} \mathbf{y}_{ig}^\top = \frac{1}{n_\bullet} \sum_{g=1}^{G} n_g \mathbf{S}_g, \tag{2}$$

where $n_\bullet = \sum_{g=1}^{G} n_g$, is an oft-used estimate of the common covariance matrix across classes. In the high-dimensional case $p > n_\bullet$ (implying $p > n_g$) the $\mathbf{S}_g$ and $\mathbf{S}_\bullet$ are singular and their inverses do not exist. Our primary interest thus lies in estimating the precision matrices $\boldsymbol{\Omega}_1 = \boldsymbol{\Sigma}_1^{-1}, \ldots, \boldsymbol{\Omega}_G = \boldsymbol{\Sigma}_G^{-1}$, as well as their commonalities and differences, when $p > n_\bullet$. We will develop a general ℓ_2-penalized ML framework to this end which we designate *targeted fused ridge estimation*.

The estimation of multiple precision matrices from high-dimensional data classes is of interest in many applications. The field of oncogenomics, for example, often deals with high-dimensional data from high-throughput experiments. Class membership may have different connotations in such settings. It may refer to certain sub-classes within a single dataset such as cancer subtypes (cancer is a very heterogeneous disease, even when present in a single organ). It may also designate different datasets or studies. Likewise, the class indicator may also refer to a conjunction of both subclass and study membership to form a two-way design of factors of interest (e.g., breast cancer subtypes present in a batch of study-specific datasets), as is often the case in oncogenomics. Our approach is thus motivated by the meta-analytic setting, where we aim for an integrative analysis in terms of simultaneously considering multiple data (sub-)classes, datasets, or both. Its desire is to borrow statistical power across classes by effectively increasing the sample size in order to improve sensitivity and specificity of discoveries.

1.1 Relation to literature and overview

There have been many proposals for estimating a single precision matrix in high-dimensional data settings. A popular approach is to amend (1) with an ℓ_1-penalty [2, 18, 47, 48]. The solution to this penalized problem is generally referred to as the *graphical lasso* and it is popular as it performs automatic model selection, i.e., the resulting estimate is sparse. It is heavily used in

Gaussian graphical modeling (GGM) as the support of a Gaussian precision matrix represents a Markov random field [24].

The ℓ_1-approach has been extended to deal with more than a single sample-group. Guo et al. [20] have proposed a parametrization of class-specific precision matrices that expresses the individual elements as a product of shared and class-specific factors. They include ℓ_1-penalties on both the shared and class-specific factors in order to jointly estimate the sparse precision matrices (representing graphical models). The penalty on the shared factors promotes a shared sparsity structure while the penalty on the class-specific factors promotes class-specific deviations from the shared sparsity structure. Danaher et al. [13] have generalized these efforts by proposing the *joint graphical lasso* which allows for various penalty structures. They study two particular choices: the *group graphical lasso* that encourages a shared sparsity structure across the class-specific precision matrices, and the *fused graphical lasso* that promotes a shared sparsity structure as well as shared precision element-values. A Bayesian approach to inferring multiple sparse precision matrices can be found in Peterson et al. [32].

While simultaneous estimation and model selection can be deemed elegant, automatic sparsity is not always an asset. It may be that one is intrinsically interested in more accurate representations of class-specific precision matrices for usage in, say, covariance-regularized regression [45] or discriminant analysis [33]. In such a situation one is not after sparse representations and one may prefer usage of a regularization method that shrinks the estimated elements of the precision matrices proportionally. In addition—when indeed considering network representations of data—the true class-specific graphical models need not be (extremely) sparse in terms of containing many zero elements. The ℓ_1-penalty is unable to retrieve the sparsity pattern when the number of truly non-null elements exceeds the available sample size [41]. In such a situation one may wish to couple a non-sparsity-inducing penalty with a post-hoc selection step allowing for probabilistic control over element selection. We therefore consider ℓ_2 or ridge-type penalization.

In Section 2 the *targeted fused ridge estimation* framework will be presented. The proposed fused ℓ_2-penalty allows for the simultaneous estimation of multiple precision matrices from high-dimensional data classes that chiefly share the same structure but that may differentiate in locations of interest. The approach is targeted in the sense that it allows for the specification of target matrices that may encode prior information. The framework is flexible and general, containing the recent work of Price et al. [33] and van Wieringen and Peeters [41] as special cases. It may be viewed as a ℓ_2-alternative to the work of Danaher et al. [13]. The method is contingent upon the selection of penalty values and target matrices, topics that are treated in Section 3. Section 4 then focuses on the graphical interpretation of precision matrices. It shows how the fused ridge precision estimates may be coupled with post-hoc support determination in order to arrive at multiple graphical models. We will refer to this coupling as the *fused graphical ridge*. This then serves as a basis for integrative

or meta-analytic network modeling. Section 5 then assesses the performance of the proposed estimator through extensive simulation experiments. Section 6 illustrates the techniques by applying it in a large scale integrative study of gene expression data of diffuse large B-cell lymphoma. The focus is then on finding common motifs and motif differences in network representations of (deregulated) molecular pathways. Section 7 concludes with a discussion.

1.2 Notation

Some additional notation must be introduced. Throughout the text and supplementary material, we use the following notation for certain matrix properties and sets: We use $\mathbf{A} \succ \mathbf{0}$ and $\mathbf{B} \succeq \mathbf{0}$ to denote symmetric positive definite (p.d.) and positive semi-definite (p.s.d.) matrices $\mathbf{A}$ and $\mathbf{B}$, respectively. By $\mathbb{R}$, $\mathbb{R}_+$, and $\mathbb{R}_{++}$ we denote the real numbers, the non-negative real numbers, and the strictly positive real numbers, respectively. In notational analogue, $\mathcal{S}^p$, $\mathcal{S}^p_+$, and $\mathcal{S}^p_{++}$ are used to denote the space of $p \times p$ real symmetric matrices, the real symmetric p.s.d. matrices, and real symmetric p.d. matrices, respectively. That is, e.g., $\mathcal{S}^p_{++} = \{\mathbf{X} \in \mathbb{R}^{p \times p} : \mathbf{X} = \mathbf{X}^\top \wedge \mathbf{X} \succ \mathbf{0}\}$. Negative subscripts similarly denote negative reals and negative definiteness. By $\mathbf{A} \geq \mathbf{B}$ and similar we denote *element-wise* relations, i.e., $(\mathbf{A})_{jq} \geq (\mathbf{B})_{jq}$ for all (j, q). Matrix subscripts will usually denote class membership, e.g., $\mathbf{A}_g$ denotes (the realization of) matrix $\mathbf{A}$ in class g. For notational brevity we will often use the shorthand $\{\mathbf{A}_g\}$ to denote the set $\{\mathbf{A}_g\}_{g=1}^G$.

The following notation is used throughout for operations: We write $\text{diag}(\mathbf{A})$ for the column vector composed of the diagonal of $\mathbf{A}$ and $\text{vec}(\mathbf{A})$ for the vectorization operator which stacks the columns of $\mathbf{A}$ on top of each other. Moreover, $\circ$ will denote the Hadamard product while $\otimes$ refers to the Kronecker product.

We will also repeatedly make use of several special matrices and functions. We let $\mathbf{I}_p$ denote the $(p \times p)$-dimensional identity matrix. Similarly, $\mathbf{J}_p$ will denote the $(p \times p)$-dimensional all-ones matrix. In addition, $\mathbf{0}$ will denote the null-matrix, the dimensions of which should be clear from the context. Lastly, $\|\cdot\|_F^2$ and $\mathbb{1}[\,\cdot\,]$ will stand for the squared Frobenius norm and the indicator function, respectively.

2 Targeted fused ridge estimation

2.1 A general penalized log-likelihood problem

Suppose G classes of $(n_g \times p)$-dimensional data exist and that the samples within each class are i.i.d. normally distributed. The log-likelihood for the data takes the following form under the additional assumption that all $n_\bullet$ observations are independent:

$$\mathcal{L}\left(\{\boldsymbol{\Omega}_g\}; \{\mathbf{S}_g\}\right) \propto \sum_g n_g \big\{\ln|\boldsymbol{\Omega}_g| - \text{tr}(\mathbf{S}_g\boldsymbol{\Omega}_g)\big\}. \tag{3}$$

We desire to obtain estimates $\{\hat{\boldsymbol{\Omega}}_g\} \in \mathcal{S}_{++}^p$ of the precision matrices for each class. Though not a requirement, we primarily consider situations in which $p > n_g$ for all g, necessitating the need for regularization. To this end, amend (3) with the *fused ridge penalty* given by

$$\begin{aligned} &f^{\mathrm{FR}}\left(\{\boldsymbol{\Omega}_g\}; \{\lambda_{g_1 g_2}\}, \{\mathbf{T}_g\}\right) \\ &\quad = \sum_g \frac{\lambda_{gg}}{2} \left\|\boldsymbol{\Omega}_g - \mathbf{T}_g\right\|_F^2 + \sum_{g_1, g_2} \frac{\lambda_{g_1 g_2}}{4} \left\|(\boldsymbol{\Omega}_{g_1} - \mathbf{T}_{g_1}) - (\boldsymbol{\Omega}_{g_2} - \mathbf{T}_{g_2})\right\|_F^2, \end{aligned} \tag{4}$$

where the $\mathbf{T}_g \in \mathcal{S}_+^p$ indicate known class-specific *target matrices* (see also Section 3.3), the $\lambda_{gg} \in \mathbb{R}_{++}$ denote class-specific *ridge penalty parameters*, and the $\lambda_{g_1 g_2} \in \mathbb{R}_+$ are pair-specific *fusion penalty parameters* subject to the requirement that $\lambda_{g_1 g_2} = \lambda_{g_2 g_1}$. All penalties can then be conveniently summarized into a non-negative symmetric matrix $\boldsymbol{\Lambda} = [\lambda_{g_1 g_2}]$ which we call the *penalty matrix*. The diagonal of $\boldsymbol{\Lambda}$ corresponds to the class-specific ridge penalties whereas off-diagonal entries are the pair-specific fusion penalties. The rationale and use of the penalty matrix is motivated further in Section 3.1. Combining (3) and (4) yields a general targeted fused ridge estimation problem:

$$\underset{\{\boldsymbol{\Omega}_g\} \in \mathcal{S}_{++}^p}{\arg\max} \left\{\mathcal{L}\left(\{\boldsymbol{\Omega}_g\}; \{\mathbf{S}_g\}\right) - f^{\mathrm{FR}}\left(\{\boldsymbol{\Omega}_g\}; \{\lambda_{g_1 g_2}\}, \{\mathbf{T}_g\}\right)\right\}. \tag{5}$$

The problem of (5) is strictly concave. Furthermore, it is worth noting that non-zero fusion penalties, $\lambda_{g_1 g_2} > 0$ for all $g_1 \neq g_2$, alone will not guarantee uniqueness when $p > n_\bullet$: In high dimensions, all ridge penalties λ_{gg} should be strictly positive to ensure identifiability. These and other properties of the estimation problem are reviewed in Section 2.2.

The problem stated in (5) is very general. We shall sometimes consider a single common ridge penalty $\lambda_{gg} = \lambda$ for all g, as well as a common fusion penalty $\lambda_{g_1 g_2} = \lambda_f$ for all class pairs $g_1 \neq g_2$ (cf., however, Section 3.1) such that $\boldsymbol{\Lambda} = \lambda \mathbf{I}_p + \lambda_f(\mathbf{J}_p - \mathbf{I}_p)$. This simplification leads to the first special case:

$$\underset{\{\boldsymbol{\Omega}_g\} \in \mathcal{S}_{++}^p}{\arg\max} \left\{\mathcal{L}\left(\{\boldsymbol{\Omega}_g\}; \{\mathbf{S}_g\}\right) - \frac{\lambda}{2} \sum_g \left\|\boldsymbol{\Omega}_g - \mathbf{T}_g\right\|_F^2 - \frac{\lambda_f}{4} \sum_{g_1, g_2} \left\|(\boldsymbol{\Omega}_{g_1} - \mathbf{T}_{g_1}) - (\boldsymbol{\Omega}_{g_2} - \mathbf{T}_{g_2})\right\|_F^2\right\}.$$

Here and analogous to (5), λ controls the rate of shrinkage of each precision $\boldsymbol{\Omega}_g$ towards the corresponding target $\mathbf{T}_g$ [41], while λ_f determines the retainment of entry-wise similarities between $(\boldsymbol{\Omega}_{g_1} - \mathbf{T}_{g_1})$ and $(\boldsymbol{\Omega}_{g_2} - \mathbf{T}_{g_2})$ for all class pairs $g_1 \neq g_2$.

When $\mathbf{T}_g = \mathbf{T}$ for all g, the problem further simplifies to

$$\underset{\{\boldsymbol{\Omega}_g\} \in \mathcal{S}_{++}^p}{\arg\max} \left\{\mathcal{L}\left(\{\boldsymbol{\Omega}_g\}; \{\mathbf{S}_g\}\right) - \frac{\lambda}{2} \sum_g \left\|\boldsymbol{\Omega}_g - \mathbf{T}\right\|_F^2 - \frac{\lambda_f}{4} \sum_{g_1, g_2} \left\|\boldsymbol{\Omega}_{g_1} - \boldsymbol{\Omega}_{g_2}\right\|_F^2\right\}, \tag{6}$$

where the targets are seen to disappear from the fusion term. Lastly, when $\mathbf{T} = \mathbf{0}$ the problem (6) reduces to its simplest form recently considered by

Price et al. [33]. Appendix A studies, in order to support an intuitive feel for the fused ridge estimation problem, its geometric interpretation in this latter context.

2.2 Estimator and properties

There is no explicit solution to (5) except for certain special cases and thus an iterative optimization procedure is needed for its general solution. As described in Section 2.3, we employ a coordinate ascent procedure which relies on the concavity of the penalized likelihood (see Lemma 5 in Appendix B.1) and repeated use of the following result, whose proof (as indeed all proofs) has been deferred to Appendix B.2:

Proposition 4
Let $\{\mathbf{T}_g\} \in \mathcal{S}_+^p$ and let $\mathbf{\Lambda} \in \mathcal{S}^G$ be a fixed penalty matrix such that $\mathbf{\Lambda} \geq \mathbf{0}$ and $\operatorname{diag}(\mathbf{\Lambda}) > \mathbf{0}$. Furthermore, assume that $\mathbf{\Omega}_g$ is p.d. and fixed for all $g \neq g_0$. The maximizing argument for class g_0 of the optimization problem (5) is then given by

$$\hat{\mathbf{\Omega}}_{g_0}(\mathbf{\Lambda}, \{\mathbf{\Omega}_g\}_{g \neq g_0}) = \left\{ \left[\bar{\lambda}_{g_0} \mathbf{I}_p + \frac{1}{4} \left(\bar{\mathbf{S}}_{g_0} - \bar{\lambda}_{g_0} \bar{\mathbf{T}}_{g_0} \right)^2 \right]^{1/2} + \frac{1}{2} \left(\bar{\mathbf{S}}_{g_0} - \bar{\lambda}_{g_0} \bar{\mathbf{T}}_{g_0} \right) \right\}^{-1} \tag{7}$$

where

$$\bar{\mathbf{S}}_{g_0} = \mathbf{S}_{g_0} - \sum_{g \neq g_0} \frac{\lambda_{g g_0}}{n_{g_0}} (\mathbf{\Omega}_g - \mathbf{T}_g), \quad \bar{\mathbf{T}}_{g_0} = \mathbf{T}_{g_0}, \quad \text{and} \quad \bar{\lambda}_{g_0} = \frac{\lambda_{g_0 \bullet}}{n_{g_0}}, \tag{8}$$

with $\lambda_{g_0 \bullet} = \sum_g \lambda_{g g_0}$ denoting the sum of the g_0th column (or row) of $\mathbf{\Lambda}$.

Remark 2.1
Defining $\bar{\mathbf{T}}_{g_0} = \mathbf{T}_{g_0}$ in Proposition 4 may be deemed redundant. However, it allows us to state equivalent alternatives to (8) without confusing notation. See Section 2.3 as well as Appendix B.2 and Section 1 of the Supplementary Material.

Remark 2.2
The target matrices from Proposition 4 may be chosen nonnegative definite. However, choosing n.d. targets may lead to ill-conditioned estimates in the limit. From a shrinkage perspective we thus prefer to choose $\{\mathbf{T}_g\} \in \mathcal{S}_{++}^p$. See Section 3.3.

Proposition 4 provides a function for updating the estimate of the g_0th class while fixing the remaining parameters. As a special case, consider the following. If all off-diagonal elements of $\boldsymbol{\Lambda}$ are zero no 'class fusion' of the estimates takes place and the maximization problem decouples into G individual, disjoint ridge estimations: See Corollary 1 in Appendix B.2. The next result summarizes some properties of (7):

Proposition 5
Consider the estimator of Proposition 4 and its accompanying assumptions. Let $\hat{\boldsymbol{\Omega}}_g \equiv \hat{\boldsymbol{\Omega}}_g\big(\boldsymbol{\Lambda}, \{\boldsymbol{\Omega}_{g'}\}_{g' \neq g}\big)$ be the precision matrix estimate of the gth class. For this estimator, the following properties hold:

i. $\hat{\boldsymbol{\Omega}}_g \succ \mathbf{0}$ for all $\lambda_{gg} \in \mathbb{R}_{++}$;

ii. $\lim_{\lambda_{gg} \to 0^+} \hat{\boldsymbol{\Omega}}_g = \mathbf{S}_g^{-1}$ if $\sum_{g' \neq g} \lambda_{gg'} = 0$ and $p \leq n_g$;

iii. $\lim_{\lambda_{gg} \to \infty^-} \hat{\boldsymbol{\Omega}}_g = \mathbf{T}_g$ if $\lambda_{gg'} < \infty$ for all $g' \neq g$;

iv. $\lim_{\lambda_{g_1 g_2} \to \infty^-} (\hat{\boldsymbol{\Omega}}_{g_1} - \mathbf{T}_{g_1}) = \lim_{\lambda_{g_1 g_2} \to \infty^-} (\hat{\boldsymbol{\Omega}}_{g_2} - \mathbf{T}_{g_2})$ if $\lambda_{g_1' g_2'} < \infty$ for all $\{g_1', g_2'\} \neq \{g_1, g_2\}$.

The first item of Proposition 5 implies that strictly positive λ_{gg} are sufficient to guarantee p.d. estimates from the ridge estimator. The second item then implies that if 'class fusion' is absent one obtains as the right-hand limit for group g the standard MLE $\mathbf{S}_g^{-1}$, whose existence is only guaranteed when $p \leq n_g$. The third item shows that the fused ridge precision estimator for class g is shrunken exactly to its target matrix when the ridge penalty tends to infinity while the fusion penalties do not. The last item shows that the precision estimators of any two classes tend to a common estimate when the fusion penalty between them tends to infinity while all remaining penalty parameters remain finite.

The attractiveness of the general estimator hinges upon the efficiency by which it can be obtained. We state a result useful in this respect before turning to our computational approach in Section 2.3:

Proposition 6
Let $\hat{\boldsymbol{\Omega}}_g \equiv \hat{\boldsymbol{\Omega}}_g\big(\boldsymbol{\Lambda}, \{\boldsymbol{\Omega}_{g'}\}_{g' \neq g}\big)$ be the precision matrix estimate (7) for the gth class and define $[\hat{\boldsymbol{\Omega}}_g]^{-1} \equiv \hat{\boldsymbol{\Sigma}}_g$. The estimate $\hat{\boldsymbol{\Omega}}_g$ can then be obtained without inversion through:

$$\begin{aligned}\hat{\boldsymbol{\Omega}}_g &= \frac{1}{\bar{\lambda}_g}\Big[\hat{\boldsymbol{\Sigma}}_g - (\bar{\mathbf{S}}_g - \bar{\lambda}_g \bar{\mathbf{T}}_g)\Big] \\ &= \frac{1}{\bar{\lambda}_g}\left\{\left[\bar{\lambda}_g \mathbf{I}_p + \frac{1}{4}\big(\bar{\mathbf{S}}_g - \bar{\lambda}_g \bar{\mathbf{T}}_{g_0}\big)^2\right]^{1/2} - \frac{1}{2}\big(\bar{\mathbf{S}}_g - \bar{\lambda}_g \bar{\mathbf{T}}_g\big)\right\}.\end{aligned}$$

2.3 Algorithm

Equation (7) allows for updating the precision estimate $\hat{\mathbf{\Omega}}_g$ of class g by plugging in the remaining $\hat{\mathbf{\Omega}}'_g$, $g' \neq g$, and assuming them fixed. Hence, from initial estimates, all precision estimates may be iteratively updated until some convergence criterion is reached. We propose a block coordinate ascent procedure to solve (5) by repeated use of the results in Proposition 4. This procedure is outlined in Algorithm 2. By the strict concavity of the problem in (5), the procedure guarantees that, contingent upon convergence, the unique maximizer is attained when considering all $\hat{\mathbf{\Omega}}_g$ jointly. Moreover, we can state the following result:

Proposition 7
The gradient ascent procedure given in Algorithm 2 will always stay within the realm of positive definite matrices $\mathcal{S}^p_{++}$.

Algorithm 2 Pseudocode for the fused ridge block coordinate ascent procedure.

Input:
Sufficient data: $(\mathbf{S}_1, n_1), \ldots, (\mathbf{S}_G, n_G)$
Penalty matrix: $\mathbf{\Lambda}$
Convergence criterion: $\varepsilon > 0$
Output:
Estimates: $\hat{\mathbf{\Omega}}_1, \ldots, \hat{\mathbf{\Omega}}_G$
procedure RIDGEP.FUSED($\mathbf{S}_1, \ldots, \mathbf{S}_G, n_1, \ldots, n_G, \mathbf{\Lambda}, \varepsilon$)
    Initialize: $\hat{\mathbf{\Omega}}_g^{(0)}$ for all g.
    for $c = 1, 2, 3, \ldots$ **do**
        for $g = 1, 2, \ldots, G$ **do**
            Update $\hat{\mathbf{\Omega}}_g^{(c)} := \hat{\mathbf{\Omega}}_g\big(\mathbf{\Lambda}, \hat{\mathbf{\Omega}}_1^{(c)}, \ldots, \hat{\mathbf{\Omega}}_{g-1}^{(c)}, \hat{\mathbf{\Omega}}_{g+1}^{(c-1)}, \ldots, \hat{\mathbf{\Omega}}_G^{(c-1)}\big)$ by (7).
        end for
        if $\max_g \left\{ \frac{\|\hat{\mathbf{\Omega}}_g^{(c)} - \hat{\mathbf{\Omega}}_g^{(c-1)}\|_F^2}{\|\hat{\mathbf{\Omega}}_g^{(c)}\|_F^2} \right\} < \varepsilon$ **then**
            return $\big(\hat{\mathbf{\Omega}}_1^{(c)}, \ldots, \hat{\mathbf{\Omega}}_G^{(c)}\big)$
        end if
    end for
end procedure

The procedure is implemented in the `rags2ridges` package within the R statistical language [34]. This implementation focuses on *stability* and *efficiency*. With regard to the former: Equivalent (in terms of the obtained estimator) alternatives to (8) can be derived that are numerically more stable for extreme

values of $\mathbf{\Lambda}$. The most apparent such alternative is:

$$\bar{\mathbf{S}}_{g_0} = \mathbf{S}_{g_0}, \quad \bar{\mathbf{T}}_{g_0} = \mathbf{T}_{g_0} + \sum_{g \neq g_0} \frac{\lambda_{gg_0}}{\lambda_{g_0\bullet}}(\mathbf{\Omega}_g - \mathbf{T}_g), \quad \text{and} \quad \bar{\lambda}_{g_0} = \frac{\lambda_{g_0\bullet}}{n_{g_0}}. \tag{9}$$

It 'updates' the target $\bar{\mathbf{T}}_g$ instead of the covariance $\bar{\mathbf{S}}_g$ and has the intuitive interpretation that the target matrix for a given class in the fused case is a combination of the actual class target matrix and the 'target corrected' estimates of remaining classes. The implementation makes use of this alternative where appropriate. See Section 1 of the Supplementary Material for details on alternative updating schemes.

Efficiency is secured through various roads. First, in certain special cases closed-form solutions to (5) exist. When appropriate, these explicit solutions are used. Moreover, these solutions may provide warm-starts for the general problem. See Section 2 of the Supplementary Material for details on estimation in these special cases. Second, the result from Proposition 6 is used, meaning that the relatively expensive operation of matrix inversion is avoided. Third, additional computational speed was achieved by implementing core operations in C++ via the R-packages `Rcpp` and `RcppArmadillo` [14, 15, 17, 38]. These efforts make analyzes with large p feasible. Throughout, we will initialize the algorithm with $\hat{\mathbf{\Omega}}_g^{(0)} = p/\operatorname{tr}(\mathbf{S}_\bullet) \cdot \mathbf{I}_p$ for all g.

3 Penalty and target selection

3.1 The penalty graph and analysis of factorial designs

Equality of all class-specific ridge penalties λ_{gg} is deemed restrictive, as is equality of all pair-specific fusion penalties $\lambda_{g_1g_2}$. In many settings, such as the analysis of factorial designs, finer control over the individual values of λ_{gg} and $\lambda_{g_1g_2}$ befits the analysis. This will be motivated by several examples of increasing complexity. In order to do so, some additional notation is developed: The penalties of $\mathbf{\Lambda}$ can be summarized by a node- and edge-weighted graph $\mathcal{P} = (W, H)$ where the vertex set W corresponds to the possible classes and the edge set H corresponds to the similarities to be retained. The weight of node $g \in W$ is given by λ_{gg} and the weight of edge $(g_1, g_2) \in H$ is then given by $\lambda_{g_1g_2}$. We refer to $\mathcal{P}$ as the *penalty graph* associated with the penalty matrix $\mathbf{\Lambda}$. The penalty graph $\mathcal{P}$ is simple and undirected as the penalty matrix is symmetric.

Example 3.1
Consider $G = 2$ classes or subtypes (ST) of diffuse large B-cell lymphoma (DLBCL) patients with tumors resembling either so-called activated B-cells (ABC) or germinal centre B-cells (GCB). Patients with the latter subtype have superior overall survival [1]. As the GCB phenotype is more common

than ABC, one might imagine a scenario where the two class sample sizes are sufficiently different such that $n_{\text{GCB}} \gg n_{\text{ABC}}$. Numeric procedures to obtain a common ridge penalty (see, e.g., Section 3.2) would then be dominated by the smaller group. Hence, choosing non-equal class ridge penalties for each group will allow for a better analysis. In such a case, the following penalty graph and matrix would be suitable:

$$\mathcal{P} = \quad \begin{array}{ccc} \text{ABC} & & \text{GCB} \\ (\lambda_{11}) & \overset{\lambda_f}{\text{———}} & (\lambda_{22}) \end{array} \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \lambda_{11} & \lambda_f \\ \lambda_f & \lambda_{22} \end{bmatrix}. \tag{10}$$

■

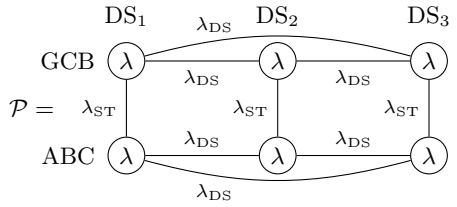
Example 3.2
Consider data from a one-way factorial design where the factor is ordinal with classes A, B, and C. For simplicity, we choose the same ridge penalty λ for each class. Say we have prior information that A is closer to B and B is closer to C than A is to C. The fusion penalty on the pairs containing the intermediate level B might then be allowed to be stronger. The following penalty graph and matrix are thus sensible:

$$\mathcal{P} = \quad \begin{array}{ccccc} \text{A} & & \text{B} & & \text{C} \\ (\lambda) & \overset{\lambda_{\text{B}}}{\text{———}} & (\lambda) & \overset{\lambda_{\text{B}}}{\text{———}} & (\lambda) \\ & \underset{\lambda_{\text{AC}}}{\frown} & & & \end{array} \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \lambda & \lambda_{\text{B}} & \lambda_{\text{AC}} \\ \lambda_{\text{B}} & \lambda & \lambda_{\text{B}} \\ \lambda_{\text{AC}} & \lambda_{\text{B}} & \lambda \end{bmatrix}. \tag{11}$$

Depending on the application, one might even omit the direct shrinkage between A and C by fixing $\lambda_{\text{AC}} = 0$. A similar penalty scheme might also be relevant if one class of the factor is an unknown mix of the remaining classes and one wishes to borrow statistical power from such a class. ■

Example 3.3
In two-way or n-way factorial designs one might wish to retain similarities in the 'direction' of each factor along with a factor-specific penalty. Consider, say, 3 oncogenic datasets (DS_1, DS_2, DS_3) regarding ABC and GCB DLBCL cancer patients. This yields a total of $G = 6$ classes of data. One choice of penalization of this 2 by 3 design is represented by the penalty graph and

matrix below:



$$\mathcal{P} = \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \lambda & \lambda_{\text{DS}} & \lambda_{\text{DS}} & \lambda_{\text{ST}} & 0 & 0 \\ \lambda_{\text{DS}} & \lambda & \lambda_{\text{DS}} & 0 & \lambda_{\text{ST}} & 0 \\ \lambda_{\text{DS}} & \lambda_{\text{DS}} & \lambda & 0 & 0 & \lambda_{\text{ST}} \\ \lambda_{\text{ST}} & 0 & 0 & \lambda & \lambda_{\text{DS}} & \lambda_{\text{DS}} \\ 0 & \lambda_{\text{ST}} & 0 & \lambda_{\text{DS}} & \lambda & \lambda_{\text{DS}} \\ 0 & 0 & \lambda_{\text{ST}} & \lambda_{\text{DS}} & \lambda_{\text{DS}} & \lambda \end{bmatrix}. \quad (12)$$

This example would favor similarities (with the same force) only between pairs sharing a common level in each factor. This finer control allows users, or the employed algorithm, to penalize differences between datasets more (or less) strongly than differences between the ABC and GCB sub-classes. This corresponds to not applying direct shrinkage of interaction effects which is of interest in some situations. ■

While the penalty graph primarily serves as an intuitive overview, it does provide some aid in the construction of the penalty matrix for multifactorial designs. For example, the construction of the penalty matrix (12) in Example 3.3 corresponds to a Cartesian graph product of two complete graphs similar to those given in (10) and (11). We state that $\mathcal{P}$ and $\boldsymbol{\Lambda}$ should be chosen carefully in conjunction with the choice of target matrices. Ideally, only strictly necessary penalization parameters (from the perspective of the desired analysis) should be introduced. Each additional penalty introduced will increase the difficulty of finding the optimal penalty values by increasing the dimension of the search-space.

3.2 Selection of penalty parameters

As the ℓ_2-penalty does not automatically induce sparsity in the estimate, it is natural to seek loss efficiency. We then use cross-validation (CV) for penalty parameter selection due to its relation to the minimization of the Kullback-Leibler divergence and its predictive accuracy stemming from its data-driven nature. Randomly divide the data of each class into $k = 1, \ldots, K$ disjoint subsets of approximately the same size. Previously, we have defined $\hat{\boldsymbol{\Omega}}_g \equiv \hat{\boldsymbol{\Omega}}_g\big(\boldsymbol{\Lambda}, \{\boldsymbol{\Omega}_{g'}\}_{g' \neq g}\big)$ to be the precision matrix estimate of the gth class. Let $\hat{\boldsymbol{\Omega}}_g^{\neg k}$ be the analogous estimate (with similar notational dependencies) for class g based on all samples not in k. Also, let $\mathbf{S}_g^k$ denote the sample covariance matrix for class g based on the data in subset k and let n_g^k denote the size of subset k in class g. The K-fold CV score for our fused regularized precision

estimate based on the fixed penalty $\boldsymbol{\Lambda}$ can then be given as:

$$\begin{aligned}\mathrm{KCV}(\boldsymbol{\Lambda}) &= \frac{1}{KG}\sum_{g=1}^{G}\sum_{k=1}^{K} n_g^k \left[-\ln|\hat{\boldsymbol{\Omega}}_g^{\neg k}| + \mathrm{tr}(\hat{\boldsymbol{\Omega}}_g^{\neg k}\mathbf{S}_g^k)\right] \\ &= -\frac{1}{KG}\sum_{g=1}^{G}\sum_{k=1}^{K} \mathcal{L}_g^k(\hat{\boldsymbol{\Omega}}_g^{\neg k}; \mathbf{S}_g^k).\end{aligned}$$

One would then choose $\boldsymbol{\Lambda}^*$ such that

$$\boldsymbol{\Lambda}^* = \arg\min_{\boldsymbol{\Lambda}} \mathrm{KCV}(\boldsymbol{\Lambda}), \quad \text{subject to:} \quad \boldsymbol{\Lambda} \geq \mathbf{0} \wedge \mathrm{diag}(\boldsymbol{\Lambda}) > \mathbf{0}. \tag{13}$$

The least biased predictive accuracy can be obtained by choosing $K = n_g$ such that $n_g^k = 1$. This would give the fused version of leave-one-out CV (LOOCV). Unfortunately, LOOCV is computationally demanding for large p and/or large n_g. We propose to select the penalties by the computationally expensive LOOCV only if adequate computational power is available. In cases where it is not, we propose two alternatives.

Our first alternative is a special version of the LOOCV scheme that significantly reduces the computational cost. The *special* LOOCV (SLOOCV) is computed much like the LOOCV. However, only the class estimate in the class of the omitted datum is updated. More specifically, the SLOOCV problem is given by:

$$\boldsymbol{\Lambda}^\diamond = \arg\min_{\boldsymbol{\Lambda}} \mathrm{SLOOCV}(\boldsymbol{\Lambda}), \quad \text{subject to:} \quad \boldsymbol{\Lambda} \geq \mathbf{0} \wedge \mathrm{diag}(\boldsymbol{\Lambda}) > \mathbf{0}, \tag{14}$$

with

$$\mathrm{SLOOCV}(\boldsymbol{\Lambda}) = -\frac{1}{n_\bullet}\sum_{g=1}^{G}\sum_{i=1}^{n_g} \mathcal{L}_g^i(\widetilde{\boldsymbol{\Omega}}_g^{\neg i}; \mathbf{S}_g^i).$$

The estimate $\widetilde{\boldsymbol{\Omega}}_g^{\neg i}$ in (14) is obtained by updating only $\hat{\boldsymbol{\Omega}}_g$ using Proposition 4. For all other $g' \neq g$, $\widetilde{\boldsymbol{\Omega}}_{g'}^{\neg i} = \hat{\boldsymbol{\Omega}}_g$. The motivation for the SLOOCV is that a single observation in a given class g does not exert heavy direct influence on the estimates in the other classes. This way the number of fused ridge estimations for each given $\boldsymbol{\Lambda}$ and each given leave-one-out sample is reduced from $n_\bullet$ to G estimations. Our second and fastest alternative is an approximation of the fused LOOCV score. This approximation can be used as an alternative to (S)LOOCV when the class sample sizes are relatively large (precisely the scenario where LOOCV is unfeasible). See Section 3 of the Supplementary Material for detailed information on this approximation.

3.3 Choice of target matrices

The target matrices $\{\mathbf{T}_g\}$ can be used to encode prior information and their choice is highly dependent on the application at hand. As they influence the

efficacy as well as the amount of bias of the estimate, it is of some importance to make a well-informed choice. Here, we describe several options of increasing level of informativeness.

In the non-fused setting, the consideration of a scalar target matrix $\mathbf{T} = \alpha\mathbf{I}_p$ for some $\alpha \in [0, \infty)$ leads to a computational benefit stemming from the property of rotation equivariance [41]: Under such targets the ridge estimator only operates on the eigenvalues of the sample covariance matrix. This benefit transfers to the fused setting for the estimator described in Proposition 4. Hence, one may consider $\mathbf{T}_g = \alpha_g\mathbf{I}_p$ with $\alpha_g \in [0, \infty)$ for each g. The limited fused ridge problem in Price et al. [33] corresponds to choosing $\alpha_g = 0$ for all g, such that a common target $\mathbf{T}_g = \mathbf{T} = \mathbf{0}$ is employed. This can be considered the least informative target possible. We generally argue against the use of the non p.d. target $\mathbf{T} = \mathbf{0}$, as it implies shrinking the class precision matrices towards the null matrix and thus towards infinite variance.

Choosing α_g to be strictly positive implies a (slightly) more informative choice. The rotation equivariance property dictates that it is sensible to choose α_g based on empirical information regarding the eigenvalues of $\mathbf{S}_g$. One such choice could be the average of the reciprocals of the non-zero eigenvalues of $\mathbf{S}_g$. A straightforward alternative would be to choose $\alpha_g = [\mathrm{tr}(\mathbf{S}_g)/p]^{-1}$. In the special case of (6) where all $\alpha_g = \alpha$ the analogous choice would be $\alpha = [\mathrm{tr}(\mathbf{S}_\bullet)/p]^{-1}$.

More informative targets would move beyond the scalar matrix. An example would be the consideration of factor-specific targets for factorial designs. Recalling Example 3.3, one might deem the dataset factor to be a 'nuisance factor'. Hence, one might choose different targets $\mathbf{T}_{\mathrm{GCB}}$ and $\mathbf{T}_{\mathrm{ABC}}$ based on training data or the pooled estimates of the GCB and ABC samples, respectively. In general, the usage of pilot training data or (pathway) database information (or both) allows for the construction of target matrices with higher specificity. We illustrate how to construct targets from database information in the DLBCL application of Section 6.

4 Fused graphical modeling

4.1 To fuse or not to fuse

As a preliminary step to downstream modeling one might consider testing the hypothesis of no class heterogeneity—and therefore the necessity of fusing—amongst the class-specific precision matrices. Effectively, one then wishes to test the null-hypothesis $H_0 : \boldsymbol{\Omega}_1 = \ldots = \boldsymbol{\Omega}_G$. Under H_0 an explicit estimator is available in which the fused penalty parameters play no role, cf. Section S2.2 of the Supplementary Material. Here we suggest a score test [6] for the evaluation of H_0 in conjunction with a way to generate its null distribution in order to assess its observational extremity.

A score test is convenient as it only requires estimation under the null

hypothesis, allowing us to exploit the availability of an explicit estimator. The score statistic equals:

$$U = -\sum_{g=1}^{G} \left(\frac{\partial \mathcal{L}(\{\boldsymbol{\Omega}_g\}; \{\mathbf{S}_g\})}{\partial \boldsymbol{\Omega}_g} \right)^{\top} \left(\frac{\partial^2 \mathcal{L}(\{\boldsymbol{\Omega}_g\}; \{\mathbf{S}_g\})}{\partial \boldsymbol{\Omega}_g \partial \boldsymbol{\Omega}_g^{\top}} \right)^{-1} \frac{\partial \mathcal{L}(\{\boldsymbol{\Omega}_g\}; \{\mathbf{S}_g\})}{\partial \boldsymbol{\Omega}_g} \Bigg|_{\boldsymbol{\Omega}_g = \hat{\boldsymbol{\Omega}}^{H_0}},$$

where $\hat{\boldsymbol{\Omega}}^{H_0}$ denotes the precision estimate under H_0 given in (S4), which holds for all classes g. The gradient can be considered in vectorized form and is readily available from (26). The Hessian of the log-likelihood equals $\partial^2 \mathcal{L} / (\partial \boldsymbol{\Omega}_g \partial \boldsymbol{\Omega}_g^{\top}) = -\boldsymbol{\Omega}_g^{-1} \otimes \boldsymbol{\Omega}_g^{-1}$. For practical purposes of evaluating the score statistic, we employ the identity $(\mathbf{A}^{\top} \otimes \mathbf{B}) \operatorname{vec}(\mathbf{C}) = \operatorname{vec}(\mathbf{BCA})$ which avoids the manipulation of $(p^2 \times p^2)$-dimensional matrices. Hence, the test statistic U is computed by

$$\hat{U} = \sum_{g=1}^{G} \operatorname{vec}(\hat{\mathbf{X}}_g)^{\top} \operatorname{vec}(\hat{\boldsymbol{\Omega}}^{H_0} \hat{\mathbf{X}}_g \hat{\boldsymbol{\Omega}}^{H_0}) = \sum_{g=1}^{G} \operatorname{tr}\big[\hat{\mathbf{X}}_g (\hat{\boldsymbol{\Omega}}^{H_0} \hat{\mathbf{X}}_g \hat{\boldsymbol{\Omega}}^{H_0})\big],$$

where $\hat{\mathbf{X}}_g = n_g \{2[(\hat{\boldsymbol{\Omega}}^{H_0})^{-1} - \mathbf{S}_g] - [(\hat{\boldsymbol{\Omega}}^{H_0})^{-1} - \mathbf{S}_g] \circ \mathbf{I}_p\}$.

The null distribution of U can be generated by permutation of the class labels: one permutes the class labels, followed by re-estimation of $\boldsymbol{\Omega}$ under H_0 and the re-calculation of the test statistic. The observed test statistic (under H_0) $\hat{U}$ is obtained from the non-permuted class labels and the regular fused estimator. The p-value is readily obtained by comparing the observed test statistic $\hat{U}$ to the null distribution obtained from the test statistic under permuted class labels. We note that the test is conditional on the choice of λ_{gg}.

4.2 Graphical modeling

A contemporary use for precision matrices is found in the reconstruction and analysis of networks through graphical modeling. Graphical models merge probability distributions of random vectors with graphs that express the conditional (in)dependencies between the constituent random variables. In the fusion setting one might think that the class precisions share a (partly) common origin (conditional independence graph) to which fusion appeals. We focus on class-specific graphs $\mathcal{G}_g = (V, E_g)$ with a finite set of vertices (or nodes) V and set of edges E_g. The vertices correspond to a collection of random variables and we consider the same set $V = \{Y_1, \ldots, Y_p\}$ of cardinality p for all classes g. That is, we consider the same p variables in all G classes. The edge set E_g is a collection of pairs of distinct vertices $(Y_j, Y_{j'})$ that are connected by an undirected edge and this collection may differ between classes. In case we assume $\{Y_1, \ldots, Y_p\} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_g)$ for all classes g we are considering multiple Gaussian graphical models.

Conditional independence between a pair of variables in the Gaussian graphical model corresponds to zero entries in the (class-specific) precision matrix. Let $\hat{\boldsymbol{\Omega}}_g$ denote a generic estimate of the precision matrix in class g. Then the

following relations hold for all pairs $\{Y_j, Y_{j'}\} \in \mathcal{V}$ with $j \neq j'$:

$$(\hat{\mathbf{\Omega}}_g)_{jj'} = \omega^{(g)}_{jj'} = 0 \iff Y_j \perp\!\!\!\perp Y_{j'} \mid V \setminus \{Y_j, Y_{j'}\} \text{ in class } g \iff (Y_j, Y_{j'}) \notin E_g.$$

Hence, determining the (in)dependence structure of the variables for class g—or equivalently the edge set E_g of $\mathcal{G}_g$—amounts to determining the support of $\hat{\mathbf{\Omega}}_g$.

4.3 Edge selection

We stress that support determination may be skipped entirely as the estimated precision matrices can be interpreted as complete (weighted) graphs. For more sparse graphical representations we resort to support determination by a local false discovery rate (lFDR) procedure [16] proposed by Schäfer and Strimmer [39]. This procedure assumes that the nonredundant off-diagonal entries of the partial correlation matrix

$$(\hat{\mathbf{P}}_g)_{jj'} = -\hat{\omega}^{(g)}_{jj'} \left(\hat{\omega}^{(g)}_{jj} \hat{\omega}^{(g)}_{j'j'} \right)^{-\frac{1}{2}}$$

follow a mixture distribution representing null and present edges. The null-distribution is known to be a scaled beta-distribution which allows for estimating the lFDR:

$$\widehat{\text{lFDR}}^{(g)}_{jj'} = P\Big((Y_j, Y_{j'}) \notin E_g \;\Big|\; (\hat{\mathbf{P}}_g)_{jj'} \Big),$$

which gives the empirical posterior probability that the edge between Y_j and $Y_{j'}$ is null in class g conditional on the observed corresponding partial correlation. The analogous probability that an edge is present can be obtained by considering $1 - \widehat{\text{lFDR}}^{(g)}_{jj'}$. See [16, 39, 41] for further details on the lFDR procedure. Our strategy will be to select for each class only those edges for which $1 - \widehat{\text{lFDR}}^{(g)}_{jj'}$ surpasses a certain threshold (see Section 6). This two-step procedure of regularization followed by subsequent support determination has the advantage that it enables probabilistic statements about the inclusion (or exclusion) of edges.

4.4 Common and differential (sub-)networks

After estimation and sparsification of the class precision matrices the identification of commonalities and differences between the graphical estimates are of natural interest. Here we consider some (summary) measures to aid such identifications. Assume in the following that multiple graphical models have been identified by the sparsified estimates $\hat{\mathbf{\Omega}}^0_1, \ldots, \hat{\mathbf{\Omega}}^0_G$ and that the corresponding graphs are denoted by $\mathcal{G}_1, \ldots, \mathcal{G}_G$.

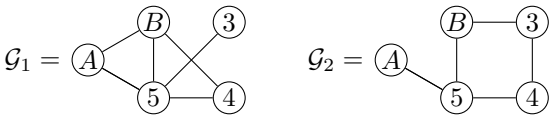
An obvious method of comparison is by pairwise graph differences or intersections. We utilize the *differential network* $\mathcal{G}_{g_1 \setminus g_2} = (V, E_{g_1} \setminus E_{g_2})$ between class g_1 and g_2 to provide an overview of edges present in one class but not the other. The *common network* $\mathcal{G}_{1 \cap 2} = (V, E_1 \cap E_2)$ is composed of the edges present in both graphs. We also define the *edge-weighted total network* of $m \leq G$ graphs $\mathcal{G}_1, \ldots, \mathcal{G}_m$ as the graph formed by the union $\mathcal{G}_{1 \cup \cdots \cup m} = (V, E_1 \cup \cdots \cup E_m)$ where the weight $w_{jj'}$ of the edge $e_{jj'}$ is given by the cardinality of the set $\{g \in \{1, \ldots, m\} : e_{jj'} \in E_g\}$. More simply, $\mathcal{G}_{1 \cup \cdots \cup m}$ is determined by summing the adjacency matrices of $\mathcal{G}_1$ to $\mathcal{G}_m$. Analogously, the *signed edge-weighted total network* takes into account the stability of the sign of an edge over the classes by summing signed adjacency matrices. Naturally, the classes can also be compared by one or more summary statistics at node-, edge-, and network-level per class, cf. [29].

We also propose the idea of 'network rewiring'. Suppose an investigator is interested in the specific interaction between genes A and B for classes g_1 and g_2. The desire is to characterize the dependency between genes A and B and determine the differences between the two classes. To do so, we suggest using the decomposition of the covariance of A and B into the individual contributions of all paths between A and B. A path z between A and B of length t_z in a graph for class g is, following Lauritzen [24], defined to be a sequence $A = v_0, \ldots, v_{t_z} = B$ of distinct vertices such that $(v_{d-1}, v_d) \in E_g$ for all $d = 1, \ldots, t_z$. The possibility of the mentioned decomposition was shown by Jones and West [22] and, in terms of $\hat{\mathbf{\Omega}}_g^0 = [\omega_{jj'}]$, can be stated as:

$$\mathrm{Cov}(A, B) = \sum_{z \in \mathcal{Z}_{AB}} (-1)^{t_z + 1} \omega_{Av_1} \omega_{v_1 v_2} \omega_{v_2 v_3} \cdots \omega_{v_{t_z - 2} v_{t_z - 1}} \omega_{v_{t_z - 1} B} \frac{|(\hat{\mathbf{\Omega}}_g^0)_{\neg P}|}{|\hat{\mathbf{\Omega}}_g^0|}, \quad (16)$$

where $\mathcal{Z}_{AB}$ is the set of all paths between A and B and $(\hat{\mathbf{\Omega}}_g^0)_{\neg P}$ denotes the matrix $\hat{\mathbf{\Omega}}_g^0$ with rows and columns corresponding to the vertices of the path z removed. Each *term* of the covariance decomposition in (16) can be interpreted as the flow of information through a given path z between A and B in $\mathcal{G}_g$. Imagine performing this decomposition for A and B in both $\hat{\mathbf{\Omega}}_{g_1}^0$ and $\hat{\mathbf{\Omega}}_{g_2}^0$. For each path, we can then identify whether it runs through the common network $\mathcal{G}_{g_1 \cap g_2}$, or utilizes the differential networks $\mathcal{G}_{g_2 \setminus g_1}, \mathcal{G}_{g_1 \setminus g_2}$ unique to the classes. The paths that pass through the differential networks can be thought of as a 'rewiring' between the groups (in particular compared to the common network). In summary, the covariance between a node pair can be separated into a component that is common and a component that is differential (or rewired).

Example 4.1
Suppose we have the following two graphs for classes $g_1 = 1$ and $g_2 = 2$:



and consider the covariance between node A and B. In $\mathcal{G}_1$ the covariance $\text{Cov}(Y_A, Y_B)$ is decomposed into contributions by the paths (A, B), $(A, 5, B)$, and $(A, 5, 4, B)$. Similarly for $\mathcal{G}_2$, the contributions are from paths $(A, 5, B)$ and $(A, 5, 4, 3, B)$. Thus $(A, 5, B)$ is the only shared path. Depending on the size of the contributions we might conclude that network 1 has some 'rewired pathways' compared to the other. This method gives a concise overview of the estimated interactions between two given genes, which genes mediate or moderate these interactions, as well as how the interaction patterns differ across the classes. In turn this might suggest candidate genes for perturbation or knock-down experiments. ■

5 Simulation study

In this section we explore and measure the performance of the fused estimator and its behavior in four different scenarios. Performance is measured primarily by the squared Frobenius loss,

$$L_F^{(g)}\big(\hat{\boldsymbol{\Omega}}_g(\boldsymbol{\Lambda}), \boldsymbol{\Omega}_g\big) = \big\|\hat{\boldsymbol{\Omega}}_g(\boldsymbol{\Lambda}) - \boldsymbol{\Omega}_g\big\|_F^2,$$

between the class precision estimate and the true population class precision matrix. However, the performance is also assessed in terms of the quadratic loss,

$$L_Q^{(g)}\big(\hat{\boldsymbol{\Omega}}_g(\boldsymbol{\Lambda}), \boldsymbol{\Omega}_g\big) = \big\|\hat{\boldsymbol{\Omega}}_g(\boldsymbol{\Lambda})\boldsymbol{\Omega}_g^{-1} - \mathbf{I}_p\big\|_F^2.$$

The risk defined as the expected loss associated with an estimator, say,

$$\mathcal{R}_F\big\{\hat{\boldsymbol{\Omega}}_g(\boldsymbol{\Lambda})\big\} = \mathbb{E}\Big[L_F^{(g)}\big(\hat{\boldsymbol{\Omega}}_g(\boldsymbol{\Lambda}), \boldsymbol{\Omega}_g\big)\Big],$$

is robustly approximated by the median loss over a repeated number of simulations and corresponding estimations.

We designed four simulation scenarios to explore the properties and performance of the fused ridge estimator and alternatives. Scenario (1) evaluates the fused ridge estimator under two choices of the penalty matrix, the non-fused ridge estimate applied individually to the classes, and the non-fused ridge estimate using the pooled covariance matrix when (1a) $\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2$ and (1b) $\boldsymbol{\Omega}_1 \neq \boldsymbol{\Omega}_2$.

Scenario (2) evaluates the fused ridge estimator under different choices of targets: (2a) $\mathbf{T}_1 = \mathbf{T}_2 = \mathbf{0}$, (2b) $\mathbf{T}_1 = \mathbf{T}_2 = \alpha\mathbf{I}_p$, and (2c) $\mathbf{T}_1 = \mathbf{T}_2 = \boldsymbol{\Omega}$. Scenario (3) evaluates the fused ridge estimator for varying network topologies and degrees of class homogeneity. Specifically, for (3a) scale-free topology and (3b) small-world topology, each with (3i) low class homogeneity and (3ii) high class homogeneity. Scenario (4) investigates the fused estimator under non-equal class sample sizes. Except for scenario 4, we make no distinction between the loss in different classes. Except for scenario 1, we use penalty matrices of the form $\boldsymbol{\Lambda} = \lambda\mathbf{I}_p + \lambda_f(\mathbf{J}_p - \mathbf{I}_p)$.

5.1 Scenario 1: Fusion versus no fusion

Scenario 1 explores the loss-efficiency of the fused estimate versus non-fused estimates as a function of the class sample size n_g for fixed p and hence for different $p/n_\bullet$ ratios. Banded population precision matrices are simulated from $G = 2$ classes. We set $p = 30$ and

$$(\boldsymbol{\Omega}_g)_{jj'} = \frac{k+1}{|j-j'|+1}\mathbb{1}\big[|j-j'| \leq k\big] \tag{17}$$

with k non-zero off-diagonal bands. The sub-scenario (1a) $\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2$ uses $k = 15$ bands whereas (1b) $\boldsymbol{\Omega}_1 \neq \boldsymbol{\Omega}_2$ uses $k = 15$ bands for $\boldsymbol{\Omega}_1$ and $k = 2$ bands for $\boldsymbol{\Omega}_2$. Hence, identical and very different population precision matrices are considered, respectively.

For $n_g = 10, 25, 70$ the loss over 20 repeated runs was computed. In each run, the optimal *unrestricted* penalty matrix $\boldsymbol{\Lambda}$ was determined by LOOCV. The losses were computed for (1i) the fused ridge estimator with an unrestricted penalty matrix, (1ii) the fused ridge estimator with a restricted penalty matrix such that $\lambda_{11} = \lambda_{22}$, (1iii) the regular non-fused ridge estimator applied separately to each class, and (1iv) the regular non-fused ridge estimator using the pooled estimate $\mathbf{S}_\bullet$. In all cases the targets $\mathbf{T}_1 = \mathbf{T}_2 = \alpha_\bullet\mathbf{I}_p$ were used with $\alpha_\bullet = p/\operatorname{tr}(\mathbf{S}_\bullet)$. The risk and quartile losses for scenario 1 are seen in the boxplots of Figure 1A.

Generally, the *unrestricted* fused estimates are found to perform at least as well as the (superior of the) *non-fused* estimates. This can be expected as the fused ridge estimate might be regarded as an interpolation between using the non-fused ridge estimator on the pooled data and within each class separately. Hence, the LOOCV procedure is thus able to capture and select the appropriate penalties both when the underlying population matrices are very similar and when they are very dissimilar. In the case of differing class population precision matrices, the *restricted* fused ridge estimator (that uses the single ridge penalty $\lambda_{11} = \lambda_{22}$) performs somewhat intermediately, indicating again the added value of the flexible penalty setup. It is unsurprising that the non-fused estimate using the pooled covariance matrix is superior in scenario 1b, where $\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2$, as it is the explicit estimator in this scenario, cf. Section S2.2 of the Supplementary Material.

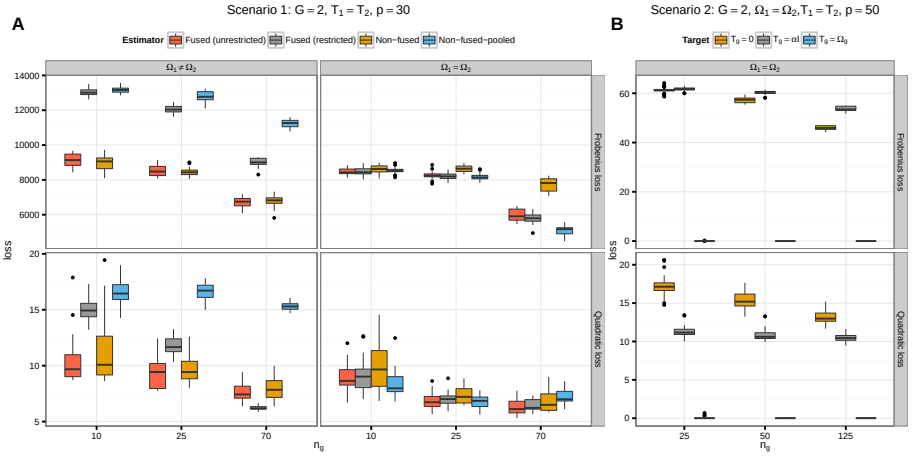


Fig. 1: A: Results for Scenario 1. The losses against the class samples size for different ridge estimators under unequal and equal class population matrices. B: Results for Scenario 2. The losses against the class sample size with different target matrices.

5.2 Scenario 2: Target versus no target

Scenario 2 investigates the added value of the targeted approach to fused precision matrix estimation compared to that of setting $\mathbf{T}_g = \mathbf{0}$ which reduces to the special-case considered by Price et al. [33]. We simulated datasets with $G = 2$ classes from banded precision matrices (as given in (17)) with $p = 50$ variables and $k = 25$ bands for varying class sample sizes n_g and target matrices $\mathbf{T}_1$ and $\mathbf{T}_2$. The performance was evaluated using (2a) $\mathbf{T}_1 = \mathbf{T}_2 = \mathbf{0}$, (2b) $\mathbf{T}_g = \alpha_\bullet \mathbf{I}_p$, as above, and (2c) the spot-on target $\mathbf{T}_1 = \mathbf{T}_2 = \boldsymbol{\Omega}$ for each of $n_g = 25, 50, 125$ class sample sizes.

As above, risks were estimated by the losses for each class over 20 simulation repetitions. The optimal penalties where determined by LOOCV with penalty matrices of the form $\boldsymbol{\Lambda} = \lambda \mathbf{I}_p + \lambda_f(\mathbf{J}_p - \mathbf{I}_p)$. The results are shown in the boxplots in Panel B of Figure 1. As expected, the spot-on target shows superior performance in terms of loss in all cases. This suggests that well-informed choices of the target can greatly improve the estimation and that the algorithm will put emphasis on the target if it reflects the truth. Such behavior is also seen analytically in the ridge estimator of Schäfer and Strimmer [39] inferred from their closed expression of the optimal penalty. We see that using the scalar target $\alpha_\bullet \mathbf{I}_p$ resuts in an as-good or lower risk in terms of the quadratic but not the Frobenius loss compared to the no-target situation.

As this scenario corresponds to the case of Price et al. [33] we performed a secondary timing benchmark of their accompanying `RidgeFusion` package compared to `rags2ridges`. We evaluated estimation time of each package on a single simulated dataset with $p = 50$, $G = 2$, and $n_1 = n_2 = 10$ using a banded matrix as before. The average estimation times over 100 model fits where 17.2

and 46.6 milliseconds for packages `rags2ridges` and `RidgeFusion`, respectively. This approximates a factor 2.71 speed-up for a single model fit. The timing was done using the package `microbenchmark` [28] and the estimates from each package were in agreement within expected numerical precision.

5.3 Scenario 3: Varying topology and class (dis)similarity

Scenario 3 investigates the fused estimator with $G = 3$ classes for (3i) high and (3ii) low class homogeneity and two different latent random graph topologies on $p = 50$ variables. The topologies are the (3a) 'small-world' and the (3b) 'scale-free' topology generated by Watts-Strogatz and Barabási graph games, respectively. The former generates topologies where all node degrees are similar while the latter game generates networks with (few) highly connected hubs. From the generated topology, we construct a latent precision matrix $\mathbf{\Psi}$ with diagonal elements set to 1 and the non-zero off-diagonal entries dictated by the network topology set to 0.1.

The two topologies are motivated as they imitate many real phenomena and processes. Small-world topologies approximate systems such as power grids, the neural network of the worm C. elegans, and the social networks of film actors [27, 43]. Conversely, scale-free topologies approximate many social networks, protein-protein interaction networks, airline networks, the world wide web, and the internet [3, 4].

We control the inter-class homogeneity using a latent inverse Wishart distribution for each class covariance matrix as considered by Bilgrau et al. [9]. That is, we let

$$\mathbf{\Sigma}_g = \mathbf{\Omega}_g^{-1} \sim \mathcal{W}_p^{-1}\Big((\nu - p - 1)\mathbf{\Phi}^{-1}, \nu\Big), \quad \nu > p + 1 \tag{18}$$

where $\mathcal{W}_p^{-1}(\mathbf{\Theta}, \nu)$ denotes an inverse Wishart distribution with scale matrix $\mathbf{\Theta}$ and ν degrees of freedom. The parametrization implies the expected value $\mathbb{E}[\mathbf{\Sigma}_g] = \mathbb{E}[\mathbf{\Omega}_g^{-1}] = \mathbf{\Phi}^{-1}$ and thus $\mathbf{\Phi}$ defines the latent expected topology. We simulate from a multivariate normal distribution as before conditional on the realized covariance $\mathbf{\Sigma}_g$. In (18), the parameter ν controls the inter-class homogeneity. Large ν imply that $\mathbf{\Omega}_1 \approx \mathbf{\Omega}_2 \approx \mathbf{\Omega}_3$ and thus a large class homogeneity. Small values of $\nu \to (p+1)^+$ imply large heterogeneity.

For the simulations, we chose (i) $\nu = 100$ and (ii) $\nu = 1000$. Again we fitted the model using both the zero target as well as the scalar matrix target described above using the reciprocal value of the mean eigenvalue, i.e., $\mathbf{T}_1 = \mathbf{T}_2 = \mathbf{T}_3 = \alpha\mathbf{I}_p$ for both $\alpha = 0$ and $\alpha = p/\operatorname{tr}(\mathbf{S}_\bullet)$. The estimation was repeated 20 times for each combination of high/low class similarity, network topology, choice of target, and class sample-size $n_1 = n_2 = n_3 = 25, 50, 125$. Panels A and B of Figure 2 show box-plots of the results.

First, the loss is seen to be dependent on the network topology, irrespective of the loss function. Second, as expected, the loss is strongly influenced by the degree of class (dis)similarity where a higher homogeneity yields a lower loss.

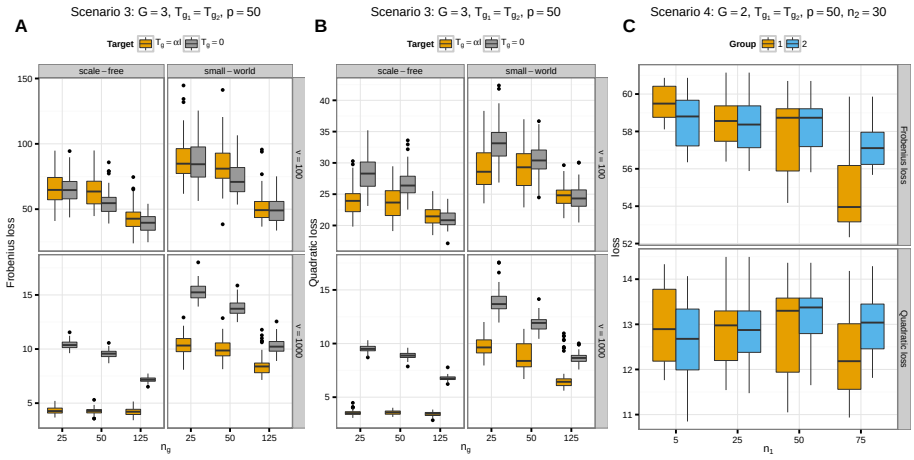


Fig. 2: A: Scenario 3. The Frobenius loss as a function of sample size under different topologies and degree of similarity. B: Scenario 3. The quadratic loss as a function of sample size under different topologies and degree of similarity. C: Scenario 4. The loss as a function of sample size of class 1 with fixed sample size for class 2.

Intuitively, this makes sense as the estimator can borrow strength across the classes and effectively increase the degrees of freedom in each class. Third, the targeted approach has a superior loss in all cases with a high class homogeneity and thus the gain in loss-efficiency is greater for the targeted approach. For low class homogeneity, the targeted approach performs comparatively to the zero target with respect to the Frobenius loss while it is seemingly better in terms of quadratic loss. Measured by quadratic loss, the targeted approach nearly always outperforms the zero target.

5.4 Scenario 4: Unequal class sizes

Scenario 4 explores the fused estimator under unequal class sample sizes. We simulated data with $k = 8$ non-zero off-diagonal bands, $G = 2$, and $p = 50$. The number of samples in class 2 was fixed at $n_2 = 30$ while the number of samples in class 1 were varied: $n_1 = 5, 25, 50, 75$. The results of the simulation are shown in Panel C of Figure 2. Note that we consider the Frobenius and quadratic loss within each class separately here.

Not surprisingly, the fused estimator performs better (for both classes) when $n_\bullet$ increases. Perhaps more surprising, there seems to be no substantial difference in quadratic loss for group n_1 and n_2 suggesting that the fusion indeed borrows strength from the larger class. A loss difference is only visible in the most extreme case where $n_1 = 5$ and $n_2 = 30$. The relative difference however is not considered large.

6 Applications

Lymphoma refers to a group of cancers that originate in specific cells of the immune system such as white blood T- or B-cells. Approximately 90% of all lymphoma cases are non-Hodgkin's lymphomas—a diverse group of blood cancers excluding Hodgkin's disease—of which the aggressive diffuse large B-cell lymphomas (DLBCL) constitutes the largest subgroup [21]. We showcase the usage of the fused ridge estimator through two analyzes of DLBCL data.

In DLBCL, there exists at least two major genetic subtypes of tumors named after their similarities in genetic expression with activated B-cells (ABC) and germinal centre B-cells (GCB). A third *umbrella* class, usually designated as Type III, contains tumors that cannot be classified as being either of the ABC or GCB subtype. Patients with tumors of GCB class show a favorable clinical prognosis compared to that of ABC. Even though the genetic subtypes have been known for more than a decade [1] and despite the appearance of refinements to the DLBCL classification system [**?**], DLBCL is still treated as a singular disease in daily clinical practice and the first differentiated treatment regimens have only recently started to appear in clinical trials [30, 36]. Many known phenotypic differences between ABC and GCB are associative, which might underline the translational inertia.fba Hence, the biological underpinnings and *functional differences* between ABC and GCB are of central interest and the motivation for the analyzes below.

Incorrect regulation of the NF-κB signaling pathway, responsible for i.a. control of cell survival, has been linked to cancer. This pathway has certain known drivers of deregulation. Aberrant interferon β production due to recurrent oncogenic mutations in the central MYD88 gene interferes with cell cycle arrest and apoptosis [46]. It also well-known that BCL2, another member of the NF-κB pathway, is deregulated in DLBCL [40]. Moreover, a deregulated NF-κB pathway is a key hallmark distinguishing the poor prognostic ABC subclass from the good prognostic GCB subclass of DLBCL [35]. Our illustrative analyzes thus focus on the *functional differences* between ABC and GCB in relation to the NF-κB pathway. Section 6.1 investigates the DLBCL classes in the context of a single dataset on the NF-κB signalling pathway. Section 6.2 analyzes multiple DLBCL NF-κB datasets with a focus on finding common motifs and motif differences in network representations of pathway-deregulation. These analyzes show the value of a fusion approach to integration. In all analyzes we take the NF-κB pathway and its constituent genes to be defined by the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [23].

6.1 Nonintegrative analysis of DLBCL subclasses

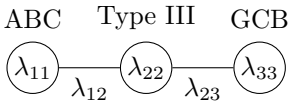
We first analyze the data from **?**], consisting of 89 DLBCL tumor samples. These samples were RMA-normalized using custom brainarray chip definition files (CDF) [12] and the R-package `affy` [19]. This preprocessing used Entrez gene identifiers (EID) by the National Center for Biotechnology Information

(NCBI), which are also used by KEGG. The usage of custom CDFs avoids the mapping problems between Affymetrix probeset IDs and KEGG. Moreover, the custom CDFs can increase the robustness and precision of the expression estimates [26, 37]. The RMA-preprocessing yielded 19,764 EIDs. Subsequently, the features were reduced to the available 82 out of the 91 EIDs present in the KEGG NF-κB pathway. The samples were then partitioned, using the DLBCL automatic classifier (DAC) by Care et al. [11], into the three classes ABC ($n_1 = 31$), III ($n_2 = 13$), and GCB ($n_3 = 45$), and gene-wise centered to have zero mean within each class.

The analysis was performed with the following settings. Target matrices for the groups were chosen to be scalar matrices with the scalar determined by the inverse of the average eigenvalue of the corresponding sample class covariance matrix, i.e.:

$$\mathbf{T}_{\text{ABC}} = \alpha_1 \mathbf{I}_p, \quad \mathbf{T}_{\text{III}} = \alpha_2 \mathbf{I}_p, \quad \mathbf{T}_{\text{GCB}} = \alpha_3 \mathbf{I}_p, \quad \text{where} \quad \alpha_g = \frac{p}{\text{tr}(\mathbf{S}_g)}.$$

These targets translate to a class-scaled 'prior' of conditional independence for all genes in NF-κB. The optimal penalties were determined by LOOCV using the penalty matrix and graph given in (19). Note that the penalty setup bears resemblance to Example 3.2. Differing class-specific ridge penalties were allowed because of considerable differences in class sample size. Direct shrinkage between ABC and GCB was disabled by fixing the corresponding pair-fusion penalty to zero. The remaining fusion penalties were free to be estimated. Usage of the Nelder-Mead optimization procedure then resulted in the optimal values $\mathbf{\Lambda}^*$ given on the right-hand side of (19) below:



$$\begin{bmatrix} \lambda_{11} & \lambda_{12} & 0 \\ \lambda_{12} & \lambda_{22} & \lambda_{23} \\ 0 & \lambda_{23} & \lambda_{33} \end{bmatrix} = \begin{bmatrix} 2 & 1.5{\cdot}10^{-3} & 0 \\ 1.5{\cdot}10^{-3} & 2.7 & 2{\cdot}10^{-3} \\ 0 & 2{\cdot}10^{-3} & 2.3 \end{bmatrix} \tag{19}$$

The ridge penalties of classes ABC and GCB are seen to be comparable in size. The small size of the Type III class leads to a relatively larger penalty to ensure a well-conditioned and stable estimate. The estimated fusion penalties are all relatively small, implying that heavy fusion is undesirable due to class-differences. The three class-specific precision matrices were estimated under $\mathbf{\Lambda}^*$ and subsequently scaled to partial correlation matrices. Panels A–C of Figure 3 visualize these partial correlation matrices. In general, the ABC and GCB classes seem to carry more signal in both the negative and positive range vis-à-vis the Type III class.

Post-hoc support determination was carried out on the partial correlation matrices using the class-wise lFDR approach of Section 4.3. The lFDR threshold was chosen conservatively to 0.99, selecting 39, 85, 34 edges for classes ABC, III, GCB, respectively. The relatively high number of edges selected for the Type III class is (at least partly) due to the difficulty of determining the mixture distribution mentioned in Section 4.3 when the overall partial correlation signal is relatively flat. Panels D–E of Figure 3 then show the conditional

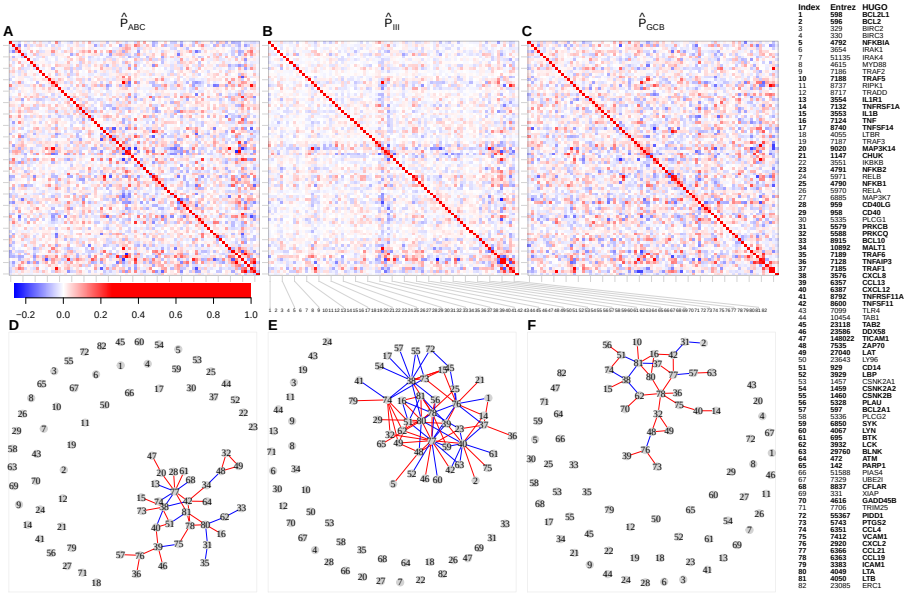


Fig. 3: *Top*: Heat maps and color key of the partial correlation matrices for the ABC (panel A), III (panel B), and GCB (panel C) classes in the NF-κB signaling pathway on the **?**] data. *Bottom*: Graphs corresponding to the sparsified precision matrices for the classes above. Red and blue edges correspond to positive and negative partial correlations, respectively. *Far right-panel*: EID key and corresponding Human Genome Organization (HUGO) Gene Nomenclature Committee (HGNC) curated gene names of the NF-κB signaling pathway genes. Genes that are connected in panels D–F are shown bold.

Table 1: The most central genes, their EID, and their plot index (I). For each class and node, the degree (with the number of positive and negative edges connected to that node in parentheses) and the betweenness centrality is shown. Only the 15 genes with the highest degrees summed over each class are shown.

			ABC		III		GCB	
	EID	I	Degree	Bet.	Degree	Bet.	Degree	Bet.
CCL21	6366	77	9 ($5^+, 4^-$)	202	17 ($9^+, 8^-$)	297	4 ($3^+, 1^-$)	106
CXCL8	3576	38	5 ($2^+, 3^-$)	126	12 ($4^+, 8^-$)	234	4 ($1^+, 3^-$)	56
CCL19	6363	78	4 ($4^+, 0^-$)	120	10 ($6^+, 4^-$)	92	6 ($6^+, 0^-$)	230
LTA	4049	80	5 ($3^+, 2^-$)	143	10 ($6^+, 4^-$)	195	3 ($3^+, 0^-$)	56
CXCL12	6387	40	3 ($2^+, 1^-$)	84	12 ($5^+, 7^-$)	187	2 ($2^+, 0^-$)	27
CXCL2	2920	76	3 ($3^+, 0^-$)	61	11 ($5^+, 6^-$)	196	3 ($2^+, 1^-$)	53
LTB	4050	81	4 ($3^+, 1^-$)	86	5 ($3^+, 2^-$)	4	6 ($3^+, 3^-$)	98
CD14	929	51	3 ($2^+, 1^-$)	20	6 ($3^+, 3^-$)	26	3 ($2^+, 1^-$)	32
CCL4	6351	74	2 ($1^+, 1^-$)	5	8 ($5^+, 3^-$)	118	2 ($1^+, 1^-$)	4
ZAP70	7535	48	3 ($2^+, 1^-$)	60	5 ($4^+, 1^-$)	51	3 ($2^+, 1^-$)	75
CCL13	6357	39	4 ($3^+, 1^-$)	119	5 ($3^+, 2^-$)	20	1 ($1^+, 0^-$)	0
TNFSF11	8600	42	5 ($4^+, 1^-$)	160	2 ($1^+, 1^-$)	0	3 ($2^+, 1^-$)	55
TNF	7124	16	1 ($1^+, 0^-$)	0	4 ($2^+, 2^-$)	2	3 ($3^+, 0^-$)	24
LAT	27040	49	2 ($2^+, 0^-$)	0	4 ($4^+, 0^-$)	16	2 ($2^+, 0^-$)	0
LCK	3932	62	2 ($0^+, 2^-$)	31	3 ($3^+, 0^-$)	10	3 ($2^+, 1^-$)	64

independence graphs corresponding to the sparsified partial correlation matrices. We note that a single connected component is identified in each class, suggesting, at least for the ABC and GCB classes, a genuine biological signal. A secondary supporting overview is provided in Table 1.

Table 1 gives the most central genes in the graphs of Panels D–E by two measures of node centrality: degree and betweenness. The node degree indicates the number of edges incident upon a particular node. The betweenness centrality indicates in how many shortest paths between vertex pairs a particular node acts as an intermediate vertex. Both measures are proxies for the importance of a feature. See, e.g., [29] for an overview of these and other centrality measures. It is seen that the CCL, CXCL, and TNF gene families are well-represented as central and connected nodes across all classes. The gene CCL21 is very central in classes ABC and III, but less so in the GCB class. From Panels D–E of Figure 3 it is seen that BCL2 and BCL2A1 are only connected in the non-ABC classes. Contrary to expectation, MYD88 is disconnected in all graphs. The genes ZAP70, LAT, and LCK found in Figure 3 and Table 1 are well-known T-cell specific genes involved in the initial T-cell receptor-mediated activation of NF-κB in T-cells [7]. From the differences in connectivity of these genes, different abundances of activated T-cells or different NF-κB activation programs for ABC/GCB might be hypothesized.

6.2 Integrative DLBCL analysis

We now expand the analysis of the previous section to show the advantages of integration by fusion. A large number of DLBCL gene expression profile (GEP) datasets is freely available at the NCBI Gene Expression Omnibus (GEO) website [5]. We obtained 11 large-scale DLBCL datasets whose GEO-accession numbers (based on various Affymetrix microarray platforms) can be found in

the first column of Table 2. One of the sets, with GEO-accession number GSE11318, is treated as a pilot/training dataset for the construction of target matrices (see below). The GSE10846 set is composed of two distinct datasets corresponding to two treatment regimens (R-CHOP and CHOP) as well as different time-periods of study. Likewise, GSE34171 is composed of three datasets corresponding to the respective microarray platforms used: HG-U133A, HG-U133B, and HG-U133 plus 2.0. As the samples on HG-U133A and HG-U133B were paired and run on *both* platforms, the (overlapping) features were averaged to form a single virtual microarray comparable to that of HG-U133 plus 2.0. Note that the **?**] data used in Section 6.1 is part of the total batch under GEO-accession number GSE56315. The sample sizes for the individual datasets vary in the range 78–495 and can also be found in Table 2. The data yield a total of 2,276 samples making this, to our knowledge, the hitherto largest integrative DLBCL study.

Similar to above, all datasets were RMA-normalized using custom brainarray CDFs and the R-package `affy`. Again, NCBI EIDs were used to avoid non-bijective gene-ID translations between the array-platforms and the KEGG database. The freely available R-package `DLBCLdata` was created to automate the download and preprocessing of the datasets in a reproducible and convenient manner. See the `DLBCLdata` documentation [8] for more information. Subsequently, the datasets were reduced to the intersecting 11,908 EIDs present on all platforms. All samples in all datasets, except for the pilot study GSE11318, were classified as either ABC, GCB, or Type III using the DAC mentioned above. The same classifier was used in all datasets to obtain a uniform classification scheme and thus maximize the comparability of the classes across datasets. Subsequently, the features were reduced to the EIDs present in the NF-κB pathway and gene-wise centered to have zero mean within each combination of DLBCL subtype and dataset. We thus have a a two-way study design—DLBCL subtypes and multiple datasets—analogous to Example 3.3. A concise overview of each of the $11 \times 3 = 33$ classes for the non-pilot data is provided in Table 2.

The target matrices were constructed from the pilot data in an attempt to use information in the directed representation $\mathcal{G}_{\text{pw}}$ of the NF-κB pathway obtained from KEGG. The directed graph represents direct and indirect causal interactions between the constituent genes. It was obtained from the KEGG database via the R-package `KEGGgraph` [49]. A target matrix was constructed for each DLCBL subtype using the pilot data and the information from the directed topology by computing node contributions using multiple linear regression models. That is, from an initial $\mathbf{T} = \mathbf{0}$, we update $\mathbf{T}$ for each node

Table 2: Overview of datasets, the defined classes, and the number of samples. In GSE31312, 28 samples were not classified with the DAC due to technical issues and hence do not appear in this table. In the pilot study GSE11318, 31 samples were primary mediastinal B-cell lymphoma and left out. Note also that the pilot dataset GSE11318 was not classified by the DAC.

	ABC		**Type III**		**GBC**		
	g	n_g	g	n_g	g	n_g	$\sum n_g$
Pilot data							
GSE11318		74		71		27	172
Dataset							
GSE56315	1	31	2	13	3	45	89
GSE19246	4	51	5	30	6	96	177
GSE12195	7	40	8	18	9	78	136
GSE22895	10	31	11	21	12	49	101
GSE31312	13	146	14	97	15	224	467
GSE10846.CHOP	16	64	17	28	18	89	181
GSE10846.RCHOP	19	75	20	42	21	116	233
GSE34171.hgu133plus2	22	23	23	15	24	52	90
GSE34171.hgu133AplusB	25	18	26	17	27	43	78
GSE22470	28	86	29	43	30	142	271
GSE4475	31	73	32	20	33	128	221
$\sum n_g$		638		344		1062	2044

$\alpha \in V(\mathcal{G}_{\text{pw}})$ through the following sequence:

$$
\begin{aligned}
T_{\alpha,\alpha} &:= T_{\alpha,\alpha} + \tfrac{1}{\sigma^2} \\
\mathbf{T}_{\text{pa}(\alpha),\alpha} &:= \mathbf{T}_{\text{pa}(\alpha),\alpha} + \tfrac{1}{\sigma^2}\boldsymbol{\beta}_{\text{pa}(\alpha)} \\
\mathbf{T}_{\alpha,\text{pa}(\alpha)} &:= \mathbf{T}_{\alpha,\text{pa}(\alpha)} + \tfrac{1}{\sigma^2}\boldsymbol{\beta}_{\text{pa}(\alpha)} \\
\mathbf{T}_{\text{pa}(\alpha),\text{pa}(\alpha)} &:= \mathbf{T}_{\text{pa}(\alpha),\text{pa}(\alpha)} + \tfrac{1}{\sigma^2}\boldsymbol{\beta}_{\text{pa}(\alpha)}\boldsymbol{\beta}_{\text{pa}(\alpha)}^\top,
\end{aligned}
$$

where $\text{pa}(\alpha)$ denotes the parents of node α in $\mathcal{G}_{\text{pw}}$, and where σ and $\boldsymbol{\beta}$ are the residual standard error and regression coefficients obtained from the linear regression of α on $\text{pa}(\alpha)$. By this scheme the target matrix represents the conditional independence structure that would result from moralizing the directed graph. If $\mathcal{G}_{\text{pw}}$ is acyclic then $\mathbf{T} \succ 0$ is guaranteed.

The penalty setup bears resemblance to Example 3.3. The Type III class is considered closer to the ABC and GCB subtypes than ABC is to GCB. Thus, the direct shrinkage between the ABC and GCB subtypes was fixed to zero. Likewise, direct shrinkage between subtype and dataset combinations was also disabled. Hence, a common ridge penalty λ, a dataset–dataset shrinkage parameter λ_{DS} and a subtype–subtype shrinkage parameter λ_{ST} were estimated. The optimal penalties were determined by SLOOCV using the penalty matrix

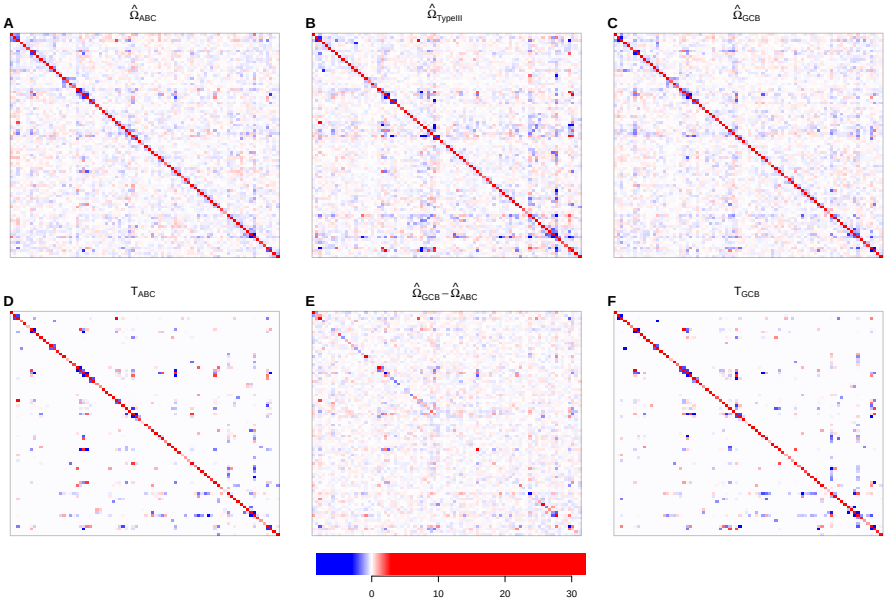
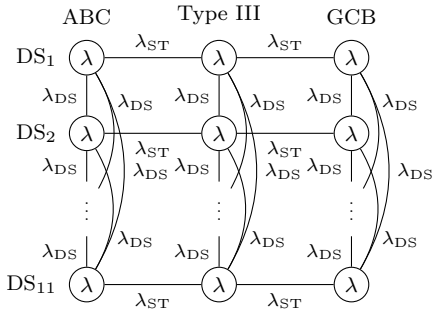


Fig. 4: Summary of the estimated precision matrices for the NF-κB pathway. *Top row*: Heat maps of the estimated precision matrices pooled across datasets for each genetic subtype. *Middle row from left to right:* The pooled target matrix for ABC, the difference between the pooled ABC and GCB estimates, and the pooled target matrix for GCB. *Bottom:* The color key for the heat maps.

and graph given in (20) below:



$$\begin{bmatrix} \lambda & \lambda_{ST} & 0 & \lambda_{DS} & 0 & 0 & \cdots & \lambda_{DS} & 0 & 0 \\ \lambda_{ST} & \lambda & \lambda_{ST} & 0 & \lambda_{DS} & 0 & \cdots & 0 & \lambda_{DS} & 0 \\ 0 & \lambda_{ST} & \lambda & 0 & 0 & \lambda_{DS} & \cdots & 0 & 0 & \lambda_{DS} \\ \lambda_{DS} & 0 & 0 & \lambda & \lambda_{ST} & 0 & \cdots & \lambda_{DS} & 0 & 0 \\ 0 & \lambda_{DS} & 0 & \lambda_{ST} & \lambda & \lambda_{ST} & \cdots & 0 & \lambda_{DS} & 0 \\ 0 & 0 & \lambda_{DS} & 0 & \lambda_{ST} & \lambda & \cdots & 0 & 0 & \lambda_{DS} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \lambda_{DS} & 0 & 0 & \lambda_{DS} & 0 & 0 & \cdots & \lambda & \lambda_{ST} & 0 \\ 0 & \lambda_{DS} & 0 & 0 & \lambda_{DS} & 0 & \cdots & \lambda_{ST} & \lambda & \lambda_{ST} \\ 0 & 0 & \lambda_{DS} & 0 & 0 & \lambda_{DS} & \cdots & 0 & \lambda_{ST} & \lambda \end{bmatrix} \tag{20}$$

The optimal penalties were found to be $\lambda^{\diamond} = 2.2$ for the ridge penalty, $\lambda^{\diamond}_{DS} = 0.0022$ for the dataset fusion penalty, and $\lambda^{\diamond}_{ST} = 6.8 \times 10^{-4}$ for the subtype fusion penalty, respectively.

To summarize and visualize the 33 class precision estimates they were pooled within DLBCL subtype. Panels A–C of Figure 4 visualizes the 3 pooled estimates as heat maps. Panels D and F visualize the constructed target matrices for the ABC and GCB subtypes, respectively. Panel E then gives the difference

between the pooled ABC and GCB estimates, indicating that they harbor differential signals to some degree. We would like to capture the commonalities and differences with a differential network representation.

The estimated class-specific precision matrices were subsequently scaled to partial correlation matrices. Each precision matrix was then sparsified using the lFDR procedure of Section 4.3. Given the class an edge was selected whenever $1 - \widehat{\text{lFDR}} \geq 0.999$. To compactly visualize the the multiple GGMs we obtained *signed edge-weighted total networks* mentioned in Section 4.4. Clearly, for inconsistent connections the weight would vary around zero, while edges that are consistently selected as positive (negative) will have a large positive (negative) weight. These meta-graphs are plotted in Figure 5. Panels A–C give the signed edge-weighted total networks for each subtype across the datasets. They show that (within DLBCL subtypes) there are a number of edges that are highly concordant across all datasets. To evaluate the greatest differences between the ABC and GCB subtypes, the signed edge-weighted total network of the latter was subtracted from the former. The resulting graph $\mathcal{G}_{\text{ABC}-\text{GCB}}$ can be found in Panel D. Edges that are more stably present in the ABC subtype are represented in orange and the edges more stably present in the GCB subtype are represented in blue. Panel F represents the graph from panel D with only those edges retained whose absolute weight exceeds 2. In a sense, the graph of panel F then represents the stable differential network. The strongest connections here should suggest places of regulatory deregulation gained or lost across the two subtypes. Interestingly, this differential network summary shows relatively large connected subgraphs suggesting differing regulatory mechanisms.

The graph in panel F of Figure 5 then conveys the added value of the integrative fusion approach. Certain members of the CCL, CXCL, and TNF gene families who were highly central in the analysis of Section 6.1 are still considered to be central here. However, it is also seen that certain genes that garnered high centrality measures in the single dataset analyzed in Section 6.1 do not behave stably *across* datasets, such as CXCL2. In addition, the integrative analysis appoints the BCL2 gene family a central role, especially in relation to the ABC subtype. This contrasts with Section 6.1, where the BCL2 gene family was not considered central and appeared to be connected mostly in the non-ABC classes. Moreover, whereas the analysis of the single dataset could not identify a signal for MYD88, the integrative analysis identifies MYD88 to be stably connected across datasets. Especially the latter two observations are in line with current knowledge on deregulation in the NF-κB pathway in DLBCL patients. Also in accordance with litterature is the known interaction of LTA with LTB seen in panel F of Figure 5 [10, 44] which here appear to be differential between ABC/GCB. Thus, borrowing information across classes enables a meta-analytic approach that can uncover information otherwise unobtainable through the analysis of single datasets.

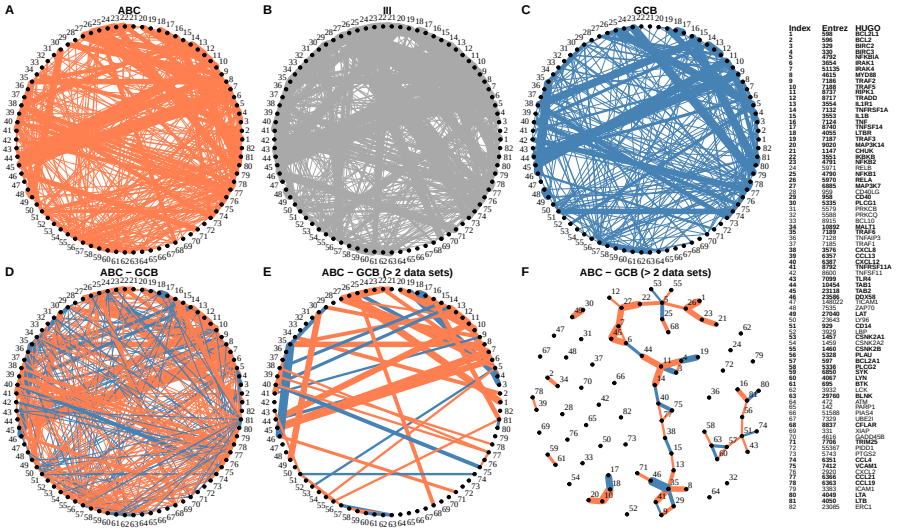


Fig. 5: Summary of estimated GGMs for the NF-κB pathway. *Panels A–C*: Graphs obtained by adding the signed adjacency matrices for each subtype across the datasets. The edge widths are drawn proportional to the absolute edge weight. *Panel D*: Graph obtained by subtracting the summarized signed adjacency matrix of GCB (panel A) from that of ABC (panel C). Edge widths are drawn proportional to the absolute weight and colored according to the sign. Orange implies edges more present in ABC and blue implies edges more present in GCB. *Panel E*: As the graph in panel D, however only edges with absolute weight > 2 are drawn. *Panel F*: As the graph in panel E, but with an alternative layout. *Far right-panel:* EID key and corresponding HGNC curated gene names of the NF-κB pathway genes. Genes that are connected in panel F are shown bold.

7 Discussion and conclusion

We considered the problem of jointly estimating multiple inverse covariance matrices from high-dimensional data consisting of distinct classes. A fused ridge estimator was proposed that generalizes previous contributions in two principal directions. First, we introduced the use of targets in fused ridge precision estimation. The targeted approach helps to stabilize the estimation procedure and allows for the incorporation of prior knowledge. It also juxtaposes itself with various alternative penalized precision matrix estimators that pull the estimates towards the edge of the parameter space, i.e., who shrink towards the non-interpretable null matrix. Second, instead of using a single ridge penalty and a single fusion penalty parameter for all classes, the approach grants the use of *class-specific* ridge penalties and *class-pair-specific* fusion penalties. This results in a flexible shrinkage framework that (i) allows for class-specific tuning, that (ii) supports analyzes when a factorial design underlies the available classes, and that (iii) supports the appropriate handling of situations where some classes are high-dimensional whilst others are low-dimensional. Targeted shrinkage and usage of a flexible penalty matrix might also benefit other procedures for precision matrix estimation such as the fused graphical lasso [13].

The targeted fused ridge estimator was combined with post-hoc support determination, which serves as a basis for integrative or meta-analytic Gaussian graphical modeling. This combination thus has applications in meta-, integrative-, and differential network analysis of multiple datasets or classes of data. This meta-approach to network analysis has multiple motivations. First, by combining data it can effectively increase the sample size in settings where samples are relatively scarce or expensive to produce. In a sense it refocuses the otherwise declining attention to obtaining a sufficient amount of data—a tendency we perceive to be untenable. Second, aggregation across multiple datasets decreases the likelihood of capturing idiosyncratic features (of individual datasets), thereby preventing over-fitting of the data.

Insightful summarization of the results is important for the feasibility of our approach to fused graphical modeling. To this end we have proposed various basic tools to summarize commonalities and differences over multiple graphs. These tools were subsequently used in a differential network analysis of the NF-κB signaling pathway in DLBCL subtypes over multiple GEP datasets. This application is not without critique, as it experiences a problem present in many GEP studies: The classification of the DLBCL subtypes (ABC and GBC) is performed on the basis of the same GEP data on which the network analysis is executed. This may be deemed methodologically undesirable. However, we justify this double use of data as (a) the pathway of interest involves a selection of genes whereas the classification uses all genes, and (b) the analysis investigates partial correlations and differential networks whereas the classification, in a sense, considers only differential expression. Furthermore, as in all large-scale genetic screenings, the analyzes should be considered 'tentative' and findings need to be validated in independent experiments. Notwithstanding,

the analyzes show that the fusion approach to network integration has merit in uncovering class-specific information on pathway deregulation. Moreover, they exemplify the exploratory *hypothesis generating* thrust of the framework we offer.

We see various inroad for further research. With regard to estimation one could think of extending the framework to incorporate a fused version of the elastic net. Mixed fusion, in the sense that one could do graphical lasso estimation with ridge fusion or ridge estimation with lasso fusion, might also be of interest. From an applied perspective the desire is to expand the toolbox for insightful (visual) summarization of commonalities and differences over multiple graphs. Moreover, it is of interest to explore improved ways for support determination. The lFDR procedure, for example, could be expanded by considering all classes jointly. Instead of applying the lFDR procedure to each class-specific precision matrix, one would then be interested in determining the proper mixture of a grand common null-distribution and multiple class-specific non-null distributions. These inroads were out of the scope of current work, but we hope to explore them elsewhere.

7.1 Software implementation

The fused ridge estimator and its accompanying estimation procedure is implemented in the `rags2ridges`-package [31] for the statistical language R. This package has many supporting functions for penalty parameter selection, graphical modeling, as well as network analysis. We will report on its full functionality elsewhere. The package is freely available from the Comprehensive R Archive Network: `http://cran.r-project.org/`.

Acknowledgements

Anders E. Bilgrau was supported by a grant from the Karen Elise Jensen Fonden, a travel grant from the Danish Cancer Society, and a visitor grant by the Dept. of Mathematics of the VU University Amsterdam. Carel F.W. Peeters received funding from the European Community's Seventh Framework Programme (FP7, 2007-2013), Research Infrastructures action, under grant agreement No. FP7-269553 (EpiRadBio project). The authors would also like to thank Karen Dybkær of the Dept. of Haematology at Aalborg University Hospital, for her help on the biological interpretations in the DLBCL application.

References

[1] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E.

Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, 2000.

[2] O. Banerjee, L. El Ghaoui, and A. D'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008.

[3] A. L. Barabási. Scale-free networks: A decade and beyond. *Science*, 325 (5939):412–413, 2009.

[4] A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[5] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva. NCBI GEO: Archive for functional genomics data sets–update. *Nucleic Acids Research*, 41(D1):D991–D995, 2013.

[6] A. K. Bera and Y. Bilias. Rao's score, Neyman's $c(\alpha)$ and Silvey's LM tests: An essay on historical developments and some new results. *Journal of Statistical Planning and Inference*, 97(1):9–44, 2001.

[7] N. Bidère, V. N. Ngo, J. Lee, C. Collins, L. Zheng, F. Wan, R. E. Davis, G. Lenz, D. E. Anderson, D. Arnoult, A. Vazquez, K. Sakai, J. Zhang, Z. Meng, T. D. Veenstra, L. M. Staudt, and M. J. Lenardo. Casein kinase 1α governs antigen-receptor-induced NF-κB activation and human lymphoma cell survival. *Nature*, 458(7234):92–96, 2009.

[8] A. E. Bilgrau and S. Falgreen. `DLBCLdata`*: Automated and Reproducible Download and Preprocessing of DLBCL Data*, 2014. URL `http://github.com/AEBilgrau/DLBCLdata`. R package version 0.9.

[9] A. E. Bilgrau, P. S. Eriksen, K. Dybkær, and M. Bøgsted. Estimation of a common covariance matrix for multiple classes with applications in meta- and discriminant analysis. *Submitted to Annals of Applied Statistics, arXiv:1503.07990*, 269553, 2015.

[10] J. L. Browning, I. D. Sizing, P. Lawton, P. R. Bourdon, P. D. Rennert, G. R. Majeau, C. M. Ambrose, C. Hession, K. Miatkowski, D. A. Griffiths, N. ek A., M. W., B. C. D., and H. P. S. Characterization of lymphotoxin-$\alpha\beta$ complexes on the surface of mouse lymphocytes. *The Journal of Immunology*, 159(7):3288–3298, 1997.

[11] M. A. Care, S. Barrans, L. Worrillow, A. Jack, D. R. Westhead, and R. M. Tooze. A microarray platform-independent classification tool for cell of origin class allows comparative analysis of gene expression in diffuse large B-cell lymphoma. *PLoS One*, 8(2):e55895, 2013.

[12] M. Dai, P. Wang, A. D. Boyd, G. Kostov, B. Athey, E. G. Jones, W. E. Bunney, R. M. Myers, T. P. Speed, H. Akil, S. J. Watson, and F. Meng. Evolving gene/transcript definitions significantly alter the interpretation of genechip data. *Nucleic Acids Research*, 33(20):e175, Jan. 2005. ISSN 1362-4962. doi: 10.1093/nar/gni179.

[13] P. Danaher, P. Wang, and D. M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society, Series B*, 76(2):373–397, 2014.

[14] D. Eddelbuettel. *Seamless R and C++ Integration with* **Rcpp**. Springer-Verlag, New York, 2013. ISBN 978-1-4614-6867-7.

[15] D. Eddelbuettel and R. François. **Rcpp**: Seamless R and C++ Integration. *Journal of Statistical Software*, 40(8), 2011.

[16] B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96:1151–1160, 2001.

[17] R. François, D. Eddelbuettel, and D. Bates. **RcppArmadillo***:* **Rcpp** *Integration for Armadillo Templated Linear Algebra Library*, 2012. URL http://CRAN.R-project.org/package=RcppArmadillo. R package version 0.3.6.1.

[18] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–41, July 2008. ISSN 1468-4357. doi: 10.1093/biostatistics/kxm045.

[19] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20(3): 307–315, 2004.

[20] Y. Guo, E. Levina, G. Michailidis, and J. Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011.

[21] International Lymphoma Study Group. A clinical evaluation of the international lymphoma study group classification of non-hodgkin's lymphoma. *Blood*, 89(11):3909–3918, June 1997. The Non-Hodgkin's Lymphoma Classification Project.

[22] B. Jones and M. West. Covariance decomposition in undirected Gaussian graphical models. *Biometrika*, 92:779–786, 2005.

[23] M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[24] S. L. Lauritzen. *Graphical models*. Clarendon Press, Oxford, 1996.

[25] H. Liu, K. Roeder, and L. Wasserman. Stability approach to regularization selection (StARS) for high dimensional graphical models. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1432–1440. Curran Associates, Inc., 2010.

[26] X. Lu and X. Zhang. The effect of GeneChip gene definitions on the microarray study of cancers. *Bioessays*, 28(7):739–46, 2006.

[27] S. Mei, X. Zhang, and M. Cao. *Power Grid Complexity*. Tsinghua University Press, Beijing and Springer-Verlag Berlin, 2011.

[28] O. Mersmann. `microbenchmark`*: Accurate Timing Functions*, 2014. URL `http://CRAN.R-project.org/package=microbenchmark`. R package version 1.4-2.

[29] M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, Oxford, 2010.

[30] G. S. Nowakowski, B. LaPlant, W. R. Macon, C. B. Reeder, J. M. Foran, G. D. Nelson, C. A. Thompson, C. E. Rivera, D. J. Inwards, I. N. Micallef, P. B. Johnston, L. F. Porrata, S. M. Ansell, R. D. Gascoyne, T. M. Habermann, and T. E. Witzig. Lenalidomide combined with R-CHOP overcomes negative prognostic impact of non-germinal center B-cell phenotype in newly diagnosed diffuse large B-cell lymphoma: A phase II study. *Journal of Clinical Oncology*, 33(3):251–257, 2015.

[31] C. F. W. Peeters, A. E. Bilgrau, and W. N. van Wieringen. `rags2ridges`*: Ridge Estimation of Precision Matrices from High-Dimensional Data*, forthcoming. R package version 2.0.

[32] C. Peterson, F. C. Stingo, and M. Vannucci. Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association*, 110(509):159–174, 2015.

[33] B. S. Price, C. J. Geyer, and A. J. Rothman. Ridge fusion in statistical learning. *Journal of Computational and Graphical Statistics*, 24(2):439–454, 2015.

[34] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

[35] M. Roschewski, L. M. Staudt, and W. H. Wilson. Diffuse large B-cell lymphoma-treatment approaches in the molecular era. *Nature Reviews Clinical Oncology*, 11(1):12–23, 2014.

[36] J. Ruan, P. Martin, R. R. Furman, S. M. Lee, K. Cheung, J. M. Vose, A. LaCasce, J. Morrison, R. Elstrom, S. Ely, A. Chadburn, E. Cesarman, M. Coleman, and J. P. Leonard. Bortezomib plus CHOP-rituximab for previously untreated diffuse large B-cell lymphoma and mantle cell lymphoma. *Journal of Clinical Oncology*, 29(6):690–697, 2011.

[37] R. Sandberg and O. Larsson. Improved precision and accuracy for microarrays using updated probe set definitions. *BMC Bioinformatics*, 8(1): 48, 2007.

[38] C. Sanderson. `Armadillo`*: An Open Source* `C++` *Linear Algebra Library for Fast Prototyping and Computationally Intensive Experiments.* Technical Report, NICTA, 2010. URL `http://arma.sourceforge.net`.

[39] J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4:art. 32, 2005.

[40] J. M. Schuetz, N. A. Johnson, R. D. Morin, D. W. Scott, K. Tan, S. Ben-Nierah, M. Boyle, G. W. Slack, M. A. Marra, J. M. Connors, A. R. Brooks-Wilson, and R. D. Gascoyne. BCL2 mutations in diffuse large B-cell lymphoma. *Leukemia*, 26(6):1383–90, 2012.

[41] W. N. van Wieringen and C. F. W. Peeters. Ridge estimation of inverse covariance matrices from high-dimensional data. Submitted to *Computational Statistics & Data Analysis, arXiv:1403.0904v3*, 2015.

[42] I. Vujačić, A. Abbruzzo, and E. Wit. A computationally fast alternative to cross-validation in penalized gaussian graphical models. *Journal of Statistical Computation and Simulation*, (ahead-of-print):1–13, 2015. doi: 10.1080/00949655.2014.992020.

[43] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[44] L. Williams-Abbott, B. N. Walter, T. C. Cheung, C. R. Goh, A. G. Porter, and C. F. Ware. The lymphotoxin-α (ltα) subunit is essential for the assembly, but not for the receptor specificity, of the membrane-anchored ltα1β2 heterotrimeric ligand. *The Journal of Biological Chemistry*, 271 (31):19451–6, 1997.

[45] D. M. Witten and R. Tibshirani. Covariance-regularized regression and classification for high-dimensional problems. *Journal of the Royal Statistical Society, Series B*, 71:615–636, 2009.

[46] Y. Yang, A. L. Shaffer, N. C. T. Emre, M. Ceribelli, M. Zhang, G. Wright, W. Xiao, J. Powell, J. Platig, H. Kohlhammer, Y. R. M., H. Zhao, Y. Yang, W. Xu, J. J. Buggy, S. Balasubramanian, L. A. Mathews, P. Shinn, R. Guha, M. Ferrer, C. Thomas, T. A. Waldmann, and L. M. Staudt. Exploiting synthetic lethality for the therapy of ABC diffuse large B cell lymphoma. *Cancer cell*, 21(6):723–737, 2012.

[47] M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94:19–35, 2007.

[48] Y. Yuan. Efficient computation of ℓ_1 regularized estimates in Gaussian graphical models. *Journal of Computational and Graphical Statistics*, 17: 809–826, 2008.

[49] J. D. Zhang and S. Wiemann. `KEGGgraph`: A graph approach to KEGG pathway in R and Bioconductor. *Bioinformatics*, 25(11):1470–1471, 2009.

A Geometric interpretation of the fused ridge penalty

Some intuition behind the fused ridge is provided by pointing to the equivalence of penalized and constrained optimization. To build this intuition we study the geometric interpretation of the fused ridge penalty in the special case of (6) with $\mathbf{T} = \mathbf{0}$. In this case $\lambda_{gg} = \lambda$ for all g, and $\lambda_{g_1 g_2} = \lambda_f$ for all $g_1 \neq g_2$. Clearly, the penalty matrix then amounts to $\mathbf{\Lambda} = \lambda \mathbf{I}_p + \lambda_f(\mathbf{J}_p - \mathbf{I}_p)$. Matters are simplified further by considering $G = 2$ classes and by focusing on a specific entry in the precision matrix, say $(\mathbf{\Omega}_g)_{jj'} = \omega^{(g)}_{jj'}$, for $g = 1, 2$. By doing so we ignore the contribution of other precision elements to the penalty. Now, the fused ridge penalty may be rewritten as:

$$\begin{aligned}
&\frac{\lambda}{2}\left(\left\|\mathbf{\Omega}_1\right\|_F^2 + \left\|\mathbf{\Omega}_2\right\|_F^2\right) + \frac{\lambda_f}{4}\sum_{g_1=1}^{2}\sum_{g_2=1}^{2}\left\|\mathbf{\Omega}_{g_1} - \mathbf{\Omega}_{g_2}\right\|_F^2 \\
&\quad = \frac{\lambda}{2}\left(\left\|\mathbf{\Omega}_1\right\|_F^2 + \left\|\mathbf{\Omega}_2\right\|_F^2\right) + \frac{\lambda_f}{2}\left\|\mathbf{\Omega}_1 - \mathbf{\Omega}_2\right\|_F^2.
\end{aligned}$$

Subsequently considering only the contribution of the $\omega^{(g)}_{jj'}$ entries implies this expression can be further reduced to:

$$\begin{aligned}
&\frac{\lambda}{2}\left[\left(\omega^{(1)}_{jj'}\right)^2 + \left(\omega^{(2)}_{jj'}\right)^2\right] + \frac{\lambda_f}{2}\left(\omega^{(1)}_{jj'} - \omega^{(2)}_{jj'}\right)^2 \\
&\quad = \frac{\lambda + \lambda_f}{2}\left[\left(\omega^{(1)}_{jj'}\right)^2 + \left(\omega^{(2)}_{jj'}\right)^2\right] - \lambda_f \omega^{(1)}_{jj'}\omega^{(2)}_{jj'}.
\end{aligned}$$

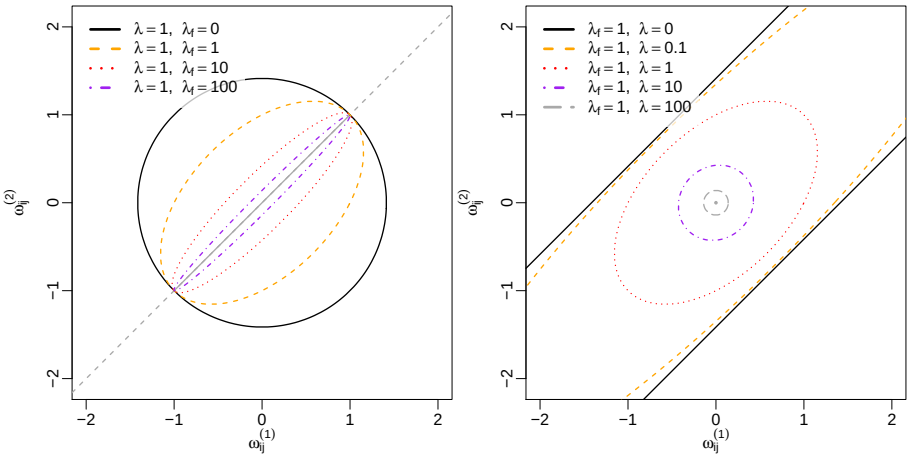


Fig. 6: Visualization of the effects of the fused ridge penalty in terms of constraints. The left panel shows the effect of λ_f for fixed λ. Here, $\lambda_f = 0$ is the regular ridge penalty. The right panel shows the effect of λ while keeping λ_f fixed.

It follows immediately that this penalty imposes constraints on the parameters $\omega^{(1)}_{jj'}$ and $\omega^{(2)}_{jj'}$, amounting to the set:

$$\left\{ \left(\omega^{(1)}_{jj'}, \omega^{(2)}_{jj'}\right) \in \mathbb{R}^2 : \frac{\lambda + \lambda_f}{2} \left[\left(\omega^{(1)}_{jj'}\right)^2 + \left(\omega^{(2)}_{jj'}\right)^2 \right] - \lambda_f \omega^{(1)}_{jj'} \omega^{(2)}_{jj'} \leq c \right\}, \tag{21}$$

for some $c \in \mathbb{R}_+$. It implies that the fused ridge penalty can be understood by the implied constraints on the parameters. Figure 6 shows the boundary of the set for selected values.

Panel 6A reveals the effect of the fused, inter-class penalty parameter λ_f (while keeping λ fixed). At $\lambda_f = 0$, the constraint coincides with the regular ridge penalty. As λ_f increases, the ellipsoid shrinks along the minor principal axis $x = -y$ with no shrinkage along $x = y$. In the limit $\lambda_f \to \infty$ the ellipsoid collapses onto the identity line. Hence, the parameters $\omega^{(1)}_{jj'}$ and $\omega^{(2)}_{jj'}$ are shrunken towards each other and while their differences vanish, their sum is not affected. Hence, the fused penalty parameter primarily shrinks the 'sum of the parameters', but also fuses them as a bound on their sizes implies a bound on their difference.

Panel 6B shows the effect of the intra-class λ penalty (while keeping λ_f fixed). When the penalty vanishes for $\lambda \to 0$ the domain becomes a degenerated ellipse (i.e. cylindrical for more than 2 classes) and parameters $\omega^{(1)}_{jj'}$ and $\omega^{(2)}_{jj'}$ may assume any value as long as their difference is less than $\sqrt{2c/\lambda_f}$. For any $\lambda > 0$, the parameter-constraint is ellipsoidal. As λ increases the ellipsoid is primarily shrunken along the principal axis formed by the identity line and along the orthogonal principal axis ($y = -x$). In the limit $\lambda \to \infty$ the ellipsoid

collapses onto the point $(0, 0)$. It is clear that the shape of the domain in (21) is only determined by the ratio of λ and λ_f.

The effect of the penalties on the domain of the obtainable estimates can be further understood by noting that the fused ridge penalty (4) can be rewritten as

$$\tilde{\lambda} \sum_{g_1,g_2} \left\|(\boldsymbol{\Omega}_{g_1}-\mathbf{T}_{g_1})+(\boldsymbol{\Omega}_{g_2}-\mathbf{T}_{g_2})\right\|_F^2+\tilde{\lambda}_f \sum_{g_1,g_2} \left\|(\boldsymbol{\Omega}_{g_1}-\mathbf{T}_{g_1})-(\boldsymbol{\Omega}_{g_2}-\mathbf{T}_{g_2})\right\|_F^2, \quad (22)$$

for some penalties $\tilde{\lambda}$ and $\tilde{\lambda}_f$. The details of this derivation can be found in Section A.1 below. The first and second summand of the rewritten penalty (22) respectively shrink the sum and difference of the parameters of the precision matrices. Their contributions thus coincide with the principal axes along which two penalty parameters shrink the domain of the parameters.

A.1 Alternative form for the fused ridge penalty

This section shows that the alternative form (22) for the ridge penalty can be written in the form (4). We again assume a common ridge penalty $\lambda_{gg} = \lambda$ and a common fusion penalty $\lambda_{g_1g_2} = \lambda_f$ for all classes and pairs thereof. To simplify the notation, let $\mathbf{A}_g = \boldsymbol{\Omega}_g - \mathbf{T}_g$. Now,

$$\begin{aligned}
&f^{\mathrm{FR}'}\left(\{\boldsymbol{\Omega}_g\}; \tilde{\lambda}, \tilde{\lambda}_f, \{\mathbf{T}_g\}\right) \\
&= \tilde{\lambda} \sum_{g_1,g_2} \left\|(\boldsymbol{\Omega}_{g_1} - \mathbf{T}_{g_1}) + (\boldsymbol{\Omega}_{g_2} - \mathbf{T}_{g_2})\right\|_F^2 + \tilde{\lambda}_f \sum_{g_1,g_2} \left\|(\boldsymbol{\Omega}_{g_1} - \mathbf{T}_{g_1}) - (\boldsymbol{\Omega}_{g_2} - \mathbf{T}_{g_2})\right\|_F^2 \\
&= \tilde{\lambda} \sum_{g_1,g_2} \left\|\mathbf{A}_{g_1} + \mathbf{A}_{g_2}\right\|_F^2 + \tilde{\lambda}_f \sum_{g_1,g_2} \left\|\mathbf{A}_{g_1} - \mathbf{A}_{g_2}\right\|_F^2 \\
&= \tilde{\lambda} \sum_{g_1,g_2} \left(\left\|\mathbf{A}_{g_1}\right\|_F^2 + \left\|\mathbf{A}_{g_2}\right\|_F^2 + 2\langle \mathbf{A}_{g_1}, \mathbf{A}_{g_2}\rangle\right) + \tilde{\lambda}_f \sum_{g_1,g_2} \left\|\mathbf{A}_{g_1} - \mathbf{A}_{g_2}\right\|_F^2 \\
&= \tilde{\lambda} \sum_{g_1,g_2} \left(2\left\|\mathbf{A}_{g_1}\right\|_F^2 + 2\left\|\mathbf{A}_{g_2}\right\|_F^2 - \left\|\mathbf{A}_{g_1} - \mathbf{A}_{g_2}\right\|_F^2\right) + \tilde{\lambda}_f \sum_{g_1,g_2} \left\|\mathbf{A}_{g_1} - \mathbf{A}_{g_2}\right\|_F^2 \\
&= 4\tilde{\lambda}G \sum_g \left\|\mathbf{A}_g\right\|_F^2 - \tilde{\lambda} \sum_{g_1,g_2} \left\|\mathbf{A}_{g_1} - \mathbf{A}_{g_2}\right\|_F^2 + \tilde{\lambda}_f \sum_{g_1,g_2} \left\|\mathbf{A}_{g_1} - \mathbf{A}_{g_2}\right\|_F^2 \\
&= 4\tilde{\lambda}G \sum_g \left\|\mathbf{A}_g\right\|_F^2 + (\tilde{\lambda}_f - \tilde{\lambda}) \sum_{g_1,g_2} \left\|\mathbf{A}_{g_1} - \mathbf{A}_{g_2}\right\|_F^2 \\
&= 4\tilde{\lambda}G \sum_g \left\|(\boldsymbol{\Omega}_g - \mathbf{T}_g)\right\|_F^2 + (\tilde{\lambda}_f - \tilde{\lambda}) \sum_{g_1,g_2} \left\|(\boldsymbol{\Omega}_{g_1} - \mathbf{T}_{g_1}) - (\boldsymbol{\Omega}_{g_2} - \mathbf{T}_{g_2})\right\|_F^2.
\end{aligned}$$

Hence, the alternative penalty (22) is also of the form (4) and thus the fused ridge of (22) is equivalent to (4) for appropriate choices of the penalties.

B Results and proofs

Section B.1 contains supporting results from other sources and results in support of Algorithm 2. Section B.2 contains proofs of the results stated in the

main text as well as additional results conducive in those proofs.

B.1 Supporting results

Lemma 3 (van Wieringen and Peeters [41])
Amend the log-likelihood (1) with the ℓ_2-penalty

$$\frac{\lambda}{2}\left\|\boldsymbol{\Omega}-\mathbf{T}\right\|_F^2,$$

with $\mathbf{T} \in \mathcal{S}_+^p$ denoting a fixed symmetric p.s.d. target matrix, and where $\lambda \in (0, \infty)$ denotes a penalty parameter. The zero gradient equation w.r.t. the precision matrix then amounts to

$$\hat{\boldsymbol{\Omega}}^{-1} - (\mathbf{S} - \lambda\mathbf{T}) - \lambda\hat{\boldsymbol{\Omega}} = \mathbf{0}, \tag{23}$$

whose solution gives a penalized ML ridge estimator of the precision matrix:

$$\hat{\boldsymbol{\Omega}}(\lambda) = \left\{ \left[\lambda\mathbf{I}_p + \frac{1}{4}(\mathbf{S} - \lambda\mathbf{T})^2 \right]^{1/2} + \frac{1}{2}(\mathbf{S} - \lambda\mathbf{T}) \right\}^{-1}.$$

Lemma 4 (van Wieringen and Peeters [41])
Consider $\hat{\boldsymbol{\Omega}}(\lambda)$ from Lemma 3 and define $[\hat{\boldsymbol{\Omega}}(\lambda)]^{-1} \equiv \hat{\boldsymbol{\Sigma}}(\lambda)$. The following identity then holds:

$$\mathbf{S} - \lambda\mathbf{T} = \hat{\boldsymbol{\Sigma}}(\lambda) - \lambda\hat{\boldsymbol{\Omega}}(\lambda).$$

Lemma 5
Let $\boldsymbol{\Lambda} \in \mathcal{S}^G$ be a matrix of fixed penalty parameters such that $\boldsymbol{\Lambda} \geq \mathbf{0}$. Moreover, let $\{\mathbf{T}_g\} \in \mathcal{S}_+^p$. Then if $\text{diag}(\boldsymbol{\Lambda}) > \mathbf{0}$, the problem of (5) is strictly concave.

Proof of Lemma 5. By $\text{diag}(\boldsymbol{\Lambda}) > \mathbf{0}$, it is clear that the fused ridge penalty (4) is strictly convex as it is a conical combination of strictly convex and convex functions. Hence, the negative fused ridge penalty is strictly concave. The log-likelihood of (3) is a conical combination of concave functions and is thus also concave. Therefore, the penalized log-likelihood is strictly concave. □

B.2 Proofs and additional results

Proof of Proposition 4. To find the maximizing argument for a specific class of the general fused ridge penalized log-likelihood problem (5) we must obtain its first-order derivative w.r.t. that class and solve the resulting zero gradient equation. To this end we first rewrite the ridge penalty (4) into a second alternative form. Using that $\mathbf{\Lambda} = \mathbf{\Lambda}^\top$, and keeping in mind the cyclic property of the trace as well as properties of $\mathbf{\Omega}_g$ and $\mathbf{T}_g$ stemming from their symmetry, we may find:

$$\begin{aligned} &f^{\mathrm{FR}''}\big(\{\mathbf{\Omega}_g\}; \mathbf{\Lambda}, \{\mathbf{T}_g\}\big) \\ &\quad= \sum_g \frac{\lambda_{gg}}{2} \big\|\mathbf{\Omega}_g - \mathbf{T}_g\big\|_F^2 + \sum_{g_1, g_2} \frac{\lambda_{g_1 g_2}}{4} \big\|(\mathbf{\Omega}_{g_1} - \mathbf{T}_{g_1}) - (\mathbf{\Omega}_{g_2} - \mathbf{T}_{g_2})\big\|_F^2 \\ &\quad= \sum_g \frac{\lambda_{g\bullet}}{2} \operatorname{tr}\Big[(\mathbf{\Omega}_g - \mathbf{T}_g)^\top (\mathbf{\Omega}_g - \mathbf{T}_g)\Big] - \sum_{\substack{g_1, g_2 \\ g_1 \neq g_2}} \frac{\lambda_{g_1 g_2}}{2} \operatorname{tr}\Big[(\mathbf{\Omega}_{g_1} - \mathbf{T}_{g_1})^\top (\mathbf{\Omega}_{g_2} - \mathbf{T}_{g_2})\Big], \end{aligned} \tag{24}$$

where $\lambda_{g\bullet} = \sum_{g'} \lambda_{gg'}$ denotes the sum over the g*th* row (or column) of $\mathbf{\Lambda}$. Taking the first-order partial derivative of (24) w.r.t. $\mathbf{\Omega}_{g_0}$ yields:

$$\begin{aligned} &\frac{\partial}{\partial \mathbf{\Omega}_{g_0}} f^{\mathrm{FR}''}\big(\{\mathbf{\Omega}_g\}; \mathbf{\Lambda}, \{\mathbf{T}_g\}\big) \\ &\quad= \lambda_{g_0\bullet} \left[2(\mathbf{\Omega}_{g_0} - \mathbf{T}_{g_0}) - (\mathbf{\Omega}_{g_0} - \mathbf{T}_{g_0}) \circ \mathbf{I}_p\right] - \sum_{g \neq g_0} \lambda_{g g_0} \left[2(\mathbf{\Omega}_g - \mathbf{T}_g) - (\mathbf{\Omega}_g - \mathbf{T}_g) \circ \mathbf{I}_p\right]. \end{aligned} \tag{25}$$

The first-order partial derivative of (3) w.r.t. $\mathbf{\Omega}_{g_0}$ results in:

$$\begin{aligned} \frac{\partial}{\partial \mathbf{\Omega}_{g_0}} \mathcal{L}(\{\mathbf{\Omega}_g\}; \{\mathbf{S}_g\}) &= \frac{\partial}{\partial \mathbf{\Omega}_{g_0}} \sum_g n_g \big\{ \ln |\mathbf{\Omega}_g| - \operatorname{tr}(\mathbf{S}_g \mathbf{\Omega}_g) \big\}, \\ &= n_{g_0} \left[2(\mathbf{\Omega}_{g_0}^{-1} - \mathbf{S}_{g_0}) - (\mathbf{\Omega}_{g_0}^{-1} - \mathbf{S}_{g_0}) \circ \mathbf{I}_p\right]. \end{aligned} \tag{26}$$

Subtracting (25) from (26) yields

$$\left[n_{g_0}(\mathbf{\Omega}_{g_0}^{-1} - \mathbf{S}_{g_0}) - \lambda_{g_0\bullet}(\mathbf{\Omega}_{g_0} - \mathbf{T}_{g_0}) + \sum_{g \neq g_0} \lambda_{g g_0}(\mathbf{\Omega}_g - \mathbf{T}_g) \right] \circ (2\mathbf{J}_p - \mathbf{I}_p), \tag{27}$$

which, clearly, is $\mathbf{0}$ only when the former factor is zero. From (27) we may then find our (conveniently scaled) zero gradient equation to be:

$$\hat{\mathbf{\Omega}}_{g_0}^{-1} - \mathbf{S}_{g_0} - \frac{\lambda_{g_0\bullet}}{n_{g_0}}(\hat{\mathbf{\Omega}}_{g_0} - \mathbf{T}_{g_0}) + \sum_{g \neq g_0} \frac{\lambda_{g g_0}}{n_{g_0}}(\mathbf{\Omega}_g - \mathbf{T}_g) = \mathbf{0}. \tag{28}$$

Now, rewrite (28) to

$$\hat{\mathbf{\Omega}}_{g_0}^{-1} - \bar{\mathbf{S}}_{g_0} - \bar{\lambda}_{g_0}(\hat{\mathbf{\Omega}}_{g_0} - \bar{\mathbf{T}}_{g_0}) = \mathbf{0}, \tag{29}$$

where $\bar{\mathbf{S}}_{g_0} = \mathbf{S}_{g_0} - \sum_{g \neq g_0} \frac{\lambda_{gg_0}}{n_{g_0}}(\boldsymbol{\Omega}_g - \mathbf{T}_g)$, $\bar{\mathbf{T}}_{g_0} = \mathbf{T}_{g_0}$, and $\bar{\lambda}_{g_0} = \lambda_{g_0\bullet}/n_{g_0}$. It can be seen that (29) is of the form (23). Lemma 3 may then be applied to obtain the solution (7). □

Corollary 1
Consider the estimator (7). Let $\hat{\boldsymbol{\Omega}}_g\big(\boldsymbol{\Lambda}, \{\boldsymbol{\Omega}_{g'}\}_{g' \neq g}\big)$ be the precision matrix estimate of the gth class. Also, let $\operatorname{diag}(\boldsymbol{\Lambda}) > \mathbf{0}$ and assume that all off-diagonal elements of $\boldsymbol{\Lambda}$ are zero. Then $\hat{\boldsymbol{\Omega}}_g\big(\boldsymbol{\Lambda}, \{\boldsymbol{\Omega}_{g'}\}_{g' \neq g}\big)$ reduces to the non-fused ridge estimate of class g:

$$\begin{aligned}\hat{\boldsymbol{\Omega}}_g\big(\boldsymbol{\Lambda}, \{\boldsymbol{\Omega}_{g'}\}_{g' \neq g}\big) &= \hat{\boldsymbol{\Omega}}_g(\lambda_{gg}) \\ &= \left\{ \left[\frac{\lambda_{gg}}{n_g}\mathbf{I}_p + \frac{1}{4}\left(\mathbf{S}_g - \frac{\lambda_{gg}}{n_g}\mathbf{T}_g\right)^2 \right]^{1/2} + \frac{1}{2}\left(\mathbf{S}_g - \frac{\lambda_{gg}}{n_g}\mathbf{T}_g\right) \right\}^{-1}. \end{aligned} \tag{30}$$

Proof of Corollary 1. The result follows directly from equations (7) and (8) by using that $\sum_{g' \neq g} \lambda_{gg'} = \sum_{g' \neq g} \lambda_{g'g} = 0$ for all g. □

Lemma 6
Let $\{\mathbf{T}_g\} \in \mathcal{S}_+^p$ and assume $\lambda_{gg} \in \mathbb{R}_{++}$ in addition to $0 \leq \lambda_{gg'} < \infty$ for all $g' \neq g$. Then

$$\lim_{\lambda_{gg} \to \infty^-} \left\| \hat{\boldsymbol{\Omega}}_g\big(\boldsymbol{\Lambda}, \{\boldsymbol{\Omega}_{g'}\}_{g' \neq g}\big) \right\|_F < \infty.$$

Proof of Lemma 6. The result is shown through proof by contradiction. Hence, suppose

$$\lim_{\lambda_{gg} \to \infty^-} \|\hat{\boldsymbol{\Omega}}_g\big(\boldsymbol{\Lambda}, \{\boldsymbol{\Omega}_{g'}\}_{g' \neq g}\big)\|_F$$

is unbounded. Let $d[\cdot]_{jj}$ denote the jth largest eigenvalue. Then, as

$$\left\| \hat{\boldsymbol{\Omega}}_g\big(\boldsymbol{\Lambda}, \{\boldsymbol{\Omega}_{g'}\}_{g' \neq g}\big) \right\|_F = \left\{ \sum_{j=1}^{p} d\left[\hat{\boldsymbol{\Omega}}_g\big(\boldsymbol{\Lambda}, \{\boldsymbol{\Omega}_{g'}\}_{g' \neq g}\big) \right]_{jj}^2 \right\}^{1/2},$$

at least one eigenvalue must tend to infinity along with λ_{gg}. Assume without loss of generality that this is only the first (and largest) eigenvalue:

$$\lim_{\lambda_{gg} \to \infty^-} d\left[\hat{\boldsymbol{\Omega}}_g\big(\boldsymbol{\Lambda}, \{\boldsymbol{\Omega}_{g'}\}_{g' \neq g}\big) \right]_{11} = \mathcal{O}(\lambda_{gg}^{\gamma}), \tag{33}$$

for some $\gamma > 0$. Now, for any λ_{gg}, the precision can be written as an eigendecomposition:

$$\hat{\boldsymbol{\Omega}}_g\big(\boldsymbol{\Lambda}, \{\boldsymbol{\Omega}_{g'}\}_{g' \neq g}\big) = d_{11}\mathbf{v}_1\mathbf{v}_1^\top + \sum_{j=2}^{p} d_{jj}\mathbf{v}_j\mathbf{v}_j^\top, \tag{34}$$

where the dependency of the eigenvalues and eigenvectors on the target matrices and penalty parameters has been suppressed (for notational brevity and clarity). It is the first summand on the right-hand side that dominates the precision for large λ_{gg}. Furthermore, this ridge ML precision estimate of the gth group satisfies, by (27), the following gradient equation:

$$n_g(\hat{\boldsymbol{\Omega}}_g^{-1} - \mathbf{S}_g) - \lambda_{gg}(\hat{\boldsymbol{\Omega}}_g - \mathbf{T}_g) - \sum_{g' \neq g} \lambda_{g'g}(\hat{\boldsymbol{\Omega}}_g - \mathbf{T}_g) + \sum_{g' \neq g} \lambda_{g'g}(\boldsymbol{\Omega}_{g'} - \mathbf{T}_{g'}) = \mathbf{0}.$$

We now make three observations: (i) From item (i) of Proposition 5 it follows that $\hat{\boldsymbol{\Omega}}_g\big(\boldsymbol{\Lambda}, \{\boldsymbol{\Omega}_{g'}\}_{g' \neq g}\big)$ is always p.d. for $\lambda_{gg} \in \mathbb{R}_{++}$. Consequently, $\lim_{\lambda_{gg} \to \infty^-} \|\hat{\boldsymbol{\Omega}}_g\big(\boldsymbol{\Lambda}, \{\boldsymbol{\Omega}_{g'}\}_{g' \neq g}\big)^{-1}\|_F < \infty$; (ii) The target matrices do not depend on λ_{gg}; and (iii) The finite $\lambda_{gg'}$ ensure that the norms of $\boldsymbol{\Omega}_{g'}$ can only exceed the norm of $\hat{\boldsymbol{\Omega}}_g$ by a function (independent of λ_{gg}) of the constant $\lambda_{gg'}$. Hence, in the limit, the norms of the $\boldsymbol{\Omega}_{g'}$ cannot exceed the norm of $\hat{\boldsymbol{\Omega}}_g$. These observations give that, as λ_{gg} tends towards infinity, the term $\lambda_{gg}(\hat{\boldsymbol{\Omega}}_g - \mathbf{T}_g)$ will dominate the gradient equation. In fact, the term $\lambda_{gg}\hat{\boldsymbol{\Omega}}_g$ will dominate as, using (33) and (34):

$$\begin{aligned}
\mathbf{0} &\approx -\lambda_{gg}(\hat{\boldsymbol{\Omega}}_g - \mathbf{T}_g) \\
&\approx -\lambda_{gg} d_{11}\mathbf{v}_1\mathbf{v}_1^\top + \lambda_{gg}\mathbf{T} \\
&\approx -\lambda_{gg}^{1+\gamma}\mathbf{v}_1\mathbf{v}_1^\top + \lambda_{gg}\mathbf{T} \\
&\approx -\lambda_{gg}^{1+\gamma}(\mathbf{v}_1\mathbf{v}_1^\top + \lambda_{gg}^{-\gamma}\mathbf{T}) \\
&\approx -\lambda_{gg}^{1+\gamma}\mathbf{v}_1\mathbf{v}_1^\top.
\end{aligned}$$

This latter statement is contradictory as it can only be true if the first eigenvalue tends to zero. This, in turn, contradicts the assumption of unboundedness (in the Frobenius norm) of the precision estimate. Hence, the fused ridge ML precision estimate must be bounded. □

Proof of Proposition 5.

(i) Note that (28) for class g may be rewritten to

$$\hat{\boldsymbol{\Omega}}_g^{-1} - \mathbf{S}_g - \frac{\lambda_{g\bullet}}{n_g}\left\{\hat{\boldsymbol{\Omega}}_g - \left[\mathbf{T}_g + \sum_{g' \neq g} \frac{\lambda_{gg'}}{\lambda_{g\bullet}}(\boldsymbol{\Omega}_{g'} - \mathbf{T}_{g'})\right]\right\} = \mathbf{0},$$

implying that (7) can be obtained under the following alternative updating scheme to (8):

$$\bar{\mathbf{S}}_g = \mathbf{S}_g, \quad \bar{\mathbf{T}}_g = \mathbf{T}_g + \sum_{g' \neq g} \frac{\lambda_{gg'}}{\lambda_{g\bullet}} (\boldsymbol{\Omega}_{g'} - \mathbf{T}_{g'}), \quad \text{and} \quad \bar{\lambda}_g = \frac{\lambda_{g\bullet}}{n_g}.$$

Now, let $d[\,\cdot\,]_{jj}$ denote the jth largest eigenvalue. Then

$$d\left\{[\hat{\boldsymbol{\Omega}}_g]^{-1}\right\}_{jj} = d\left[\frac{1}{2}(\mathbf{S}_g - \bar{\lambda}_g \bar{\mathbf{T}}_g)\right]_{jj} + \sqrt{\left\{d\left[\frac{1}{2}(\mathbf{S}_g - \bar{\lambda}_g \bar{\mathbf{T}}_g)\right]_{jj}\right\}^2 + \bar{\lambda}_g} > 0,$$

when $\bar{\lambda}_g > 0$. As $\bar{\lambda}_g = \sum_{g'}(\lambda_{g'g}/n_g)$ and as $\lambda_{g'g}$ may be 0 for all $g' \neq g$, $\hat{\boldsymbol{\Omega}}_g$ is guaranteed to be p.d. whenever $\lambda_{gg} \in \mathbb{R}_{++}$.

(ii) Note that $\sum_{g' \neq g} \lambda_{gg'} = \sum_{g' \neq g} \lambda_{g'g} = 0$ implies that $\hat{\boldsymbol{\Omega}}_g$ reduces to the non-fused class estimate (30) by way of Corollary 1. The stated right-hand limit is then immediate by using $\lambda_{gg} = 0$ in (30). Under the distributional assumptions this limit exists with probability 1 when $p \leq n_g$.

(iii) Consider the zero gradient equation (28) for the gth class. Multiply it by $n_g/\lambda_{g\bullet}$ to factor out the dominant term:

$$\frac{n_g}{\lambda_{g\bullet}} \hat{\boldsymbol{\Omega}}_g^{-1} - \frac{n_g}{\lambda_{g\bullet}} \mathbf{S}_g - (\hat{\boldsymbol{\Omega}}_g - \mathbf{T}_g) + \sum_{g' \neq g} \frac{\lambda_{g'g}}{\lambda_{g\bullet}} (\boldsymbol{\Omega}_{g'} - \mathbf{T}_{g'}) = \mathbf{0}. \tag{38}$$

When $\lambda_{gg} \to \infty^-$, $\lambda_{g\bullet} = \sum_{g'} \lambda_{gg'} \to \infty^-$, implying that the first two terms of (38) vanish. Under the assumption that $\lambda_{gg'} < \infty$ for all $g' \neq g$ we have that $\lambda_{g'g}/\lambda_{g\bullet} \to 0$ when $\lambda_{gg} \to \infty^-$ for all $g' \neq g$. Thus, all terms of the sum also vanish as Lemma 6 implies that the $\boldsymbol{\Omega}_{g'}$ are all bounded. Hence, when $\lambda_{gg} \to \infty^-$ and $\lambda_{gg'} < \infty$ for all $g' \neq g$, the zero gradient equation reduces to $\hat{\boldsymbol{\Omega}}_g - \mathbf{T}_g = \mathbf{0}$, implying the stated left-hand limit.

(iv) The proof strategy follows the proof of item iii. Multiply the zero gradient equation (28) for the g_1th class with $n_{g_1}/\lambda_{g_1 g_2}$ to obtain:

$$\frac{n_{g_1}}{\lambda_{g_1 g_2}} \hat{\boldsymbol{\Omega}}_{g_1}^{-1} - \frac{n_{g_1}}{\lambda_{g_1 g_2}} \mathbf{S}_{g_1} - \frac{\lambda_{g_1\bullet}}{\lambda_{g_1 g_2}} (\hat{\boldsymbol{\Omega}}_{g_1} - \mathbf{T}_{g_1}) + \sum_{g' \neq g_1} \frac{\lambda_{g' g_1}}{\lambda_{g_1 g_2}} (\boldsymbol{\Omega}_{g'} - \mathbf{T}_{g'}) = \mathbf{0}. \tag{39}$$

The first two terms are immediately seen to vanish when $\lambda_{g_1 g_2} \to \infty^-$. Under the assumption that all penalties except $\lambda_{g_1 g_2}$ are finite, we have that $\lambda_{g_1\bullet}/\lambda_{g_1 g_2} \to 1$ for $\lambda_{g_1 g_2} \to \infty^-$. Similarly, all elements of the sum term in (39) vanish except the element where $g' = g_2$. Hence, when $\lambda_{g_1 g_2} \to \infty^-$ and when $\lambda_{g_1' g_2'} < \infty$ for all $\{g_1', g_2'\} \neq \{g_1, g_2\}$, the zero gradient equation for class g_1 reduces to:

$$-(\hat{\boldsymbol{\Omega}}_{g_1} - \mathbf{T}_{g_1}) + (\boldsymbol{\Omega}_{g_2} - \mathbf{T}_{g_2}) = \mathbf{0}. \tag{40}$$

Conversely, by multiplying the zero gradient equation (28) for the g_2th class with $n_{g_2}/\lambda_{g_1 g_2}$ one obtains, through the same development as above, that the zero gradient equation for class g_2 reduces to the $\hat{\boldsymbol{\Omega}}_{g_2}$-analogy of equation (40). The result (40) then immediately implies the stated limiting result. □

Corollary 2
Consider item iv of Proposition 5. When, in addition, $\mathbf{T}_{g_1} = \mathbf{T}_{g_2}$, we have that

$$\lim_{\lambda_{g_1 g_2} \to \infty^-} (\hat{\mathbf{\Omega}}_{g_1} - \mathbf{T}_{g_1}) = \lim_{\lambda_{g_1 g_2} \to \infty^-} (\hat{\mathbf{\Omega}}_{g_2} - \mathbf{T}_{g_2}) \qquad \Longrightarrow \qquad \hat{\mathbf{\Omega}}_{g_1} = \hat{\mathbf{\Omega}}_{g_2}.$$

Proof of Corollary 2. The implication follows directly by using $\mathbf{T}_{g_1} = \mathbf{T}_{g_2}$ in (40). □

Proof of Proposition 6. The result follows from Proposition 4 and Lemma 4.□

Proof of Proposition 7. Note that line 8 of Algorithm 2 implies that the initializing estimates are p.d. Moreover, regardless of the value of the fused penalties (in the feasible domain), the estimate in line 11 of Algorithm 2 is p.d. as a consequence of Proposition 5. □

Supplementary Materials

This is the supplementary material for the paper *'Targeted Fused Ridge Estimation of Inverse Covariance Matrices from Multiple High-Dimensional Data Classes.'*

S1 Alternative fused ridge solutions

This section derives two equivalent, in terms of (7), but alternative updating schemes to (8). The motivation for the exploration of these alternative recursive estimators is twofold. First, alternative recursions can exhibit differing numerical (in)stability for extreme values of the penalty matrix $\boldsymbol{\Lambda} = [\lambda_{g_1 g_2}]$. Second, they provide additional intuition and understanding of the targeted fused ridge estimator.

The general strategy to finding the alternatives is to rewrite the gradient equation (28) into the non-fused form (29), which we will repeat here:

$$\hat{\boldsymbol{\Omega}}_{g_0}^{-1} - \bar{\mathbf{S}}_{g_0} - \bar{\lambda}_{g_0}(\hat{\boldsymbol{\Omega}}_{g_0} - \bar{\mathbf{T}}_{g_0}) = \mathbf{0}, \tag{S1}$$

where $\bar{\lambda}_{g_0}$, $\bar{\mathbf{T}}_{g_0}$, and $\bar{\mathbf{S}}_{g_0}$ do not depend on $\hat{\boldsymbol{\Omega}}_{g_0}$. Note that an explicit closed-form solution to (S1) exists in the form of (7).

S1.1 First alternative

The first alternative scheme is straightforward. Rewrite (28) to:

$$\begin{aligned}\mathbf{0} &= n_{g_0}\hat{\boldsymbol{\Omega}}_{g_0}^{-1} - n_{g_0}\mathbf{S}_{g_0} - \lambda_{g_0\bullet}(\hat{\boldsymbol{\Omega}}_{g_0} - \mathbf{T}_{g_0}) + \sum_{g \neq g_0} \lambda_{g g_0}(\boldsymbol{\Omega}_g - \mathbf{T}_g) \\ &= n_{g_0}\hat{\boldsymbol{\Omega}}_{g_0}^{-1} - n_{g_0}\mathbf{S}_{g_0} - \lambda_{g_0\bullet}\left\{\hat{\boldsymbol{\Omega}}_{g_0} - \left[\mathbf{T}_{g_0} + \sum_{g \neq g_0} \frac{\lambda_{g g_0}}{\lambda_{g_0\bullet}}(\boldsymbol{\Omega}_g - \mathbf{T}_g)\right]\right\},\end{aligned} \tag{S2}$$

where $\lambda_{g_0\bullet} = \sum_g \lambda_{g g_0}$. In terms of (S1), we thus have the updating scheme given in equation (9). As stated in the main text, it has the intuitive interpretation that a fused class target is used which is a combination of the class-specific target and the 'target corrected' estimates of remaining classes.

S1.2 Second alternative

We now derive a second alternative recursion scheme. Add and subtract $\lambda_{g_0\bullet}\sum_{g\neq g_0}\lambda_{gg_0}\boldsymbol{\Omega}_g$, to (S2) and rewrite such that:

$$\begin{aligned}
\mathbf{0} &= n_{g_0}\hat{\boldsymbol{\Omega}}_{g_0}^{-1} - n_{g_0}\mathbf{S}_{g_0} - \lambda_{g_0\bullet}(\hat{\boldsymbol{\Omega}}_{g_0} - \mathbf{T}_{g_0}) + \lambda_{g_0\bullet}\sum_{g\neq g_0}\lambda_{gg_0}\boldsymbol{\Omega}_g \\
&\qquad + \sum_{g\neq g_0}\lambda_{gg_0}(\boldsymbol{\Omega}_g - \mathbf{T}_g) - \lambda_{g_0\bullet}\sum_{g\neq g_0}\lambda_{gg_0}\boldsymbol{\Omega}_g \\
&= n_{g_0}\hat{\boldsymbol{\Omega}}_{g_0}^{-1} - n_{g_0}\mathbf{S}_{g_0} - \lambda_{g_0\bullet}\left[\hat{\boldsymbol{\Omega}}_{g_0} - \left(\mathbf{T}_{g_0} + \sum_{g\neq g_0}\lambda_{gg_0}\boldsymbol{\Omega}_g\right)\right] \\
&\qquad + \sum_{g\neq g_0}\lambda_{gg_0}\boldsymbol{\Omega}_g - \sum_{g\neq g_0}\lambda_{gg_0}\mathbf{T}_g - \lambda_{g_0\bullet}\sum_{g\neq g_0}\lambda_{gg_0}\boldsymbol{\Omega}_g \\
&= n_{g_0}\hat{\boldsymbol{\Omega}}_{g_0}^{-1} - n_{g_0}\mathbf{S}_{g_0} - \lambda_{g_0\bullet}\left[\hat{\boldsymbol{\Omega}}_{g_0} - \left(\mathbf{T}_{g_0} + \sum_{g\neq g_0}\lambda_{gg_0}\boldsymbol{\Omega}_g\right)\right] \\
&\qquad - \sum_{g\neq g_0}\lambda_{gg_0}\mathbf{T}_g - (\lambda_{g_0\bullet}-1)\sum_{g\neq g_0}\lambda_{gg_0}\boldsymbol{\Omega}_g \\
&= n_{g_0}\hat{\boldsymbol{\Omega}}_{g_0}^{-1} - n_{g_0}\left[\mathbf{S}_{g_0} + \frac{\lambda_{g_0\bullet}-1}{n_{g_0}}\sum_{g\neq g_0}\lambda_{gg_0}\boldsymbol{\Omega}_g + \sum_{g\neq g_0}\frac{\lambda_{gg_0}}{n_{g_0}}\mathbf{T}_g\right] \\
&\qquad - \lambda_{g_0\bullet}\left[\hat{\boldsymbol{\Omega}}_{g_0} - \left(\mathbf{T}_{g_0} + \sum_{g\neq g_0}\lambda_{gg_0}\boldsymbol{\Omega}_g\right)\right].
\end{aligned}$$

Dividing by n_{g_0} gives

$$\begin{aligned}
\mathbf{0} = \hat{\boldsymbol{\Omega}}_{g_0}^{-1} &- \left[\mathbf{S}_{g_0} + \frac{\lambda_{g_0\bullet}-1}{n_{g_0}}\sum_{g\neq g_0}\lambda_{gg_0}\boldsymbol{\Omega}_g + \sum_{g\neq g_0}\frac{\lambda_{gg_0}}{n_{g_0}}\mathbf{T}_g\right] \\
&\qquad - \frac{\lambda_{g_0\bullet}}{n_{g_0}}\left[\hat{\boldsymbol{\Omega}}_{g_0} - \left(\mathbf{T}_{g_0} + \sum_{g\neq g_0}\lambda_{gg_0}\boldsymbol{\Omega}_g\right)\right],
\end{aligned}$$

which brings the expression to the desired form (S1) with the updating scheme

$$\begin{aligned}
\bar{\mathbf{S}}_{g_0} &= \mathbf{S}_{g_0} + \frac{\lambda_{g_0\bullet}-1}{n_{g_0}}\sum_{g\neq g_0}\lambda_{gg_0}\boldsymbol{\Omega}_g + \sum_{g\neq g_0}\frac{\lambda_{gg_0}}{n_{g_0}}\mathbf{T}_g, \\
\bar{\mathbf{T}}_{g_0} &= \mathbf{T}_{g_0} + \sum_{g\neq g_0}\lambda_{gg_0}\boldsymbol{\Omega}_g, \quad \text{and} \quad \bar{\lambda}_{g_0} = \frac{\lambda_{g_0\bullet}}{n_{g_0}}.
\end{aligned}$$

Again, a solution for $\hat{\boldsymbol{\Omega}}_{g_0}$ with fixed $\boldsymbol{\Omega}_g$ for all $g \neq g_0$, is available through Lemma 3 [41] and is given in (7).

S1.3 Motivation

Though seemingly more complicated, these alternative updating schemes can be numerically more stable for extreme penalties. In both alternatives, we see

that $\bar{\mathbf{S}}_{g_0}$ is p.s.d. for (nearly) all very large and very small penalties. Likewise, $\bar{\mathbf{T}}_{g_0}$ is always positive definite. Compare the alternative expressions to the updating scheme given by (8) which can be seen to be numerically unstable for very large penalties: For very large λ_{gg} or $\lambda_{g_1g_2}$ the $\bar{\mathbf{S}}_{g_0}$ in (8) may be a matrix with numerically extreme values. This implies ill-conditioning and numerical instability under finite computer precision. On the other hand, 'updating' the target matrix will generally lead to updates for which the resulting estimator is not rotationally equivariant. This implies a reduction in computational speed.

S2 Estimation in special cases

Here we explore scenarios for which we arrive at explicit targeted fused ridge estimators. These explicit solutions further insight into the behavior of the general estimator and they can provide computational speed-ups in certain situations. Three special cases are covered:

I. $\lambda_{gg'} = 0$ for all $g \neq g'$ or equivalently $\sum_{g'} \lambda_{gg'} = \lambda_{g\bullet} = \lambda_{gg}$ for all g;

II. $\boldsymbol{\Omega}_1 = \cdots = \boldsymbol{\Omega}_G$ and $\mathbf{T}_g = \mathbf{T}$ for all g;

III. $\mathbf{T}_g = \mathbf{T}$ for all g, $\lambda_{gg} = \lambda$ for all g, $\lambda_{g_1g_2} = \lambda_f$ for all $g_1 \neq g_2$, and $\lambda_f \to \infty^-$.

S2.1 Special case I

When $\sum_{g'} \lambda_{gg'} = \lambda_{g\bullet} = \lambda_{gg}$ for all g, we have that $\sum_{g' \neq g} \lambda_{gg'} = \sum_{g' \neq g} \lambda_{g'g} = 0$ for all g. Hence, all fusion penalties are zero. The zero gradient equation (28) for class g then no longer hinges upon information from the remaining classes g'. The targeted fused precision estimate for class g then reduces to (30) of Corollary 1. This case thus coincides, as expected, with obtaining G decoupled non-fused ridge precision estimates. A special case that results in the same estimates occurs when considering $\lambda_{g_1g_2} = \lambda_f$ for all $g_1 \neq g_2$ and λ_f is taken to be 0.

S2.2 Special case II

Suppose $\boldsymbol{\Omega}_g = \boldsymbol{\Omega}$ and $\mathbf{T}_g = \mathbf{T}$ for all g. Consequently, the fusion penalty term vanishes irrespective of the values of the $\lambda_{g_1g_2}$, $g_1 \neq g_2$. The zero gradient equation (28) then reduces to

$$\mathbf{0} = n_g \hat{\boldsymbol{\Omega}}^{-1} - n_g \mathbf{S}_g - \lambda_{gg}(\hat{\boldsymbol{\Omega}} - \mathbf{T}),$$

for each class g. Adding all G equations implies:

$$\begin{aligned}\mathbf{0} &= \sum_{g=1}^{G} n_g \hat{\boldsymbol{\Omega}}^{-1} - \sum_{g=1}^{G} n_g \mathbf{S}_g - \left(\sum_{g=1}^{G} \lambda_{gg}\right)(\hat{\boldsymbol{\Omega}} - \mathbf{T}) \\ &= n_\bullet \hat{\boldsymbol{\Omega}}^{-1} - n_\bullet \mathbf{S}_\bullet - \operatorname{tr}(\boldsymbol{\Lambda})(\hat{\boldsymbol{\Omega}} - \mathbf{T}) \\ &= \hat{\boldsymbol{\Omega}}^{-1} - \left[\mathbf{S}_\bullet - \frac{\operatorname{tr}(\boldsymbol{\Lambda})}{n_\bullet}\mathbf{T}\right] - \frac{\operatorname{tr}(\boldsymbol{\Lambda})}{n_\bullet}\hat{\boldsymbol{\Omega}}. \end{aligned} \tag{S3}$$

We recognize that (S3) is of the form (23). Lemma 3 may then be directly applied to obtain the solution:

$$\hat{\boldsymbol{\Omega}}(\boldsymbol{\Lambda}) = \left\{\left[\lambda^* \mathbf{I}_p + \frac{1}{4}(\mathbf{S}_\bullet - \lambda^* \mathbf{T})^2\right]^{1/2} + \frac{1}{2}(\mathbf{S}_\bullet - \lambda^* \mathbf{T})\right\}^{-1}, \tag{S4}$$

where $\lambda^* = \operatorname{tr}(\boldsymbol{\Lambda})/n_\bullet$. Hence, this second special case gives a non-fused penalized estimate that uses the pooled covariance matrix. It can be interpreted as an averaged penalized estimator. It is of importance in testing equality of the class precision matrices (see Section 4.1 of the main text).

S2.3 Special case III

Suppose that $\mathbf{T}_g = \mathbf{T}$ for all g, that $\lambda_{gg} = \lambda$ for all g, and that $\lambda_{g_1 g_2} = \lambda_f$ for all $g_1 \neq g_2$. The main optimization problem then reduces to (6). Clearly, for $\lambda_f \to \infty^-$ the fused penalty

$$f^{\mathrm{FR}}(\{\boldsymbol{\Omega}_\mathbf{g}\}; \lambda, \lambda_f, \mathbf{T}) = \frac{\lambda}{2}\sum_g \left\|\boldsymbol{\Omega}_g - \mathbf{T}\right\|_F^2 + \frac{\lambda_f}{4}\sum_{g_1, g_2} \left\|(\boldsymbol{\Omega}_{g_1} - \boldsymbol{\Omega}_{g_2})\right\|_F^2$$

is minimized when $\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2 = \cdots = \boldsymbol{\Omega}_G$. This is also implied, more rigorously, by Corollary 2. Hence, the problem reduces to the special case of section S2.2 considered above. The solution to the penalized ML problem when $\lambda_f = \infty$ is then given by (S4) where $\operatorname{tr}(\boldsymbol{\Lambda})$ now implies $G\lambda$.

S3 Fused Kullback-Leibler approximate cross-validation

S3.1 Motivation

In ℓ_1-penalized estimation of the precision matrix, penalty selection implies (graphical) model selection: Regularization results in automatic selection of conditional dependencies. One then seeks to select an optimal value for the penalty parameter in terms of model selection consistency. To this end, the Bayesian information criterion (BIC), the extended BIC (EBIC), and the stability approach to regularization selection (StARS) are appropriate [25]. The

(fused) ℓ_2-penalty will not directly induce sparsity in precision matrix estimates. Hence, in ℓ_2-penalized problems it is natural to choose the penalty parameters on the basis of efficiency loss. Of interest are then estimators of the Kullback-Leibler (KL) divergence, such as LOOCV, generalized approximate cross-validation (GACV), and Akaike's information criterion (AIC). While superior in terms of predictive accuracy due to its data-driven nature, the LOOCV is computationally very expensive. Vujačić et al. [42] proposed a KL-based CV loss with superior performance to both AIC and GACV. The proposed method has closed-form solutions and thus provides a fast approximation to LOOCV. Here, we extend this method to provide a computationally friendly approximation of the fused LOOCV score.

S3.2 Formulation

Following Vujačić et al. [42], we now restate the KL approximation to LOOCV in the fused ridge setting. Let the true precision matrix for class g be denoted by $\boldsymbol{\Omega}_g$. Its estimate, shorthanded by $\hat{\boldsymbol{\Omega}}_g$ can be obtained through Algorithm 2. The KL divergence between the multivariate normal distributions $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Omega}_g^{-1})$ and $\mathcal{N}_p(\mathbf{0}, \hat{\boldsymbol{\Omega}}_g^{-1})$ can be shown to be:

$$\mathrm{KL}(\boldsymbol{\Omega}_g, \hat{\boldsymbol{\Omega}}_g) = \frac{1}{2}\Big\{\mathrm{tr}(\boldsymbol{\Omega}_g^{-1}\hat{\boldsymbol{\Omega}}_g) - \ln|\boldsymbol{\Omega}_g^{-1}\hat{\boldsymbol{\Omega}}_g| - p\Big\}.$$

For each g we wish to minimize this divergence. In the fused case we therefore consider the *fused Kullback-Leibler* (FKL) divergence which, motivated by the LOOCV score, is taken to be a weighted average of KL divergences:

$$\begin{aligned}&\mathrm{FKL}(\{\boldsymbol{\Omega}_g\}, \{\hat{\boldsymbol{\Omega}}_g\}) \\ &\quad= \frac{1}{n_\bullet}\sum_{g=1}^{G} n_g \,\mathrm{KL}(\boldsymbol{\Omega}_g, \hat{\boldsymbol{\Omega}}_g) = \frac{1}{n_\bullet}\sum_{g=1}^{G} \frac{n_g}{2}\Big\{\mathrm{tr}(\boldsymbol{\Omega}_g^{-1}\hat{\boldsymbol{\Omega}}_g) - \ln|\boldsymbol{\Omega}_g^{-1}\hat{\boldsymbol{\Omega}}_g| - p\Big\}.\end{aligned} \tag{S5}$$

The FKL divergence (S5) can, using the likelihood (3), be rewritten as

$$\begin{aligned}\mathrm{FKL} &= -\frac{1}{n_\bullet}\mathcal{L}\big(\{\hat{\boldsymbol{\Omega}}_g\}; \{\mathbf{S}_g\}\big) + \mathrm{bias}, \quad \text{where} \\ \mathrm{bias} &= \frac{1}{2n_\bullet}\sum_{g=1}^{G} n_g \,\mathrm{tr}\big[\hat{\boldsymbol{\Omega}}_g(\boldsymbol{\Omega}_g^{-1} - \mathbf{S}_g)\big],\end{aligned}$$

and where the equality holds up to the addition of a constant. It is clear that the bias term depends on the unknown true precision matrices and thus needs to be estimated. The fused analogue to the proposal of Vujačić et al. [42], called the *fused Kullback-Leibler approximate cross-validation* score or simply *approximate fused LOOCV* score, then is

$$\widehat{\mathrm{FKL}}(\boldsymbol{\Lambda}) = -\frac{1}{n_\bullet}\mathcal{L}\big(\{\hat{\boldsymbol{\Omega}}_g\}; \{\mathbf{S}_g\}\big) + \widehat{\mathrm{bias}}, \tag{S6}$$

with

$$\widehat{\text{bias}} = \frac{1}{2n_\bullet} \sum_{g=1}^{G} \sum_{i=1}^{n_g} \Big\{ \mathbf{y}_{ig}^\top (\hat{\boldsymbol{\Omega}}_g^2 - \hat{\boldsymbol{\Omega}}_g) \mathbf{y}_{ig} + \bar{\lambda}_g \mathbf{y}_{ig}^\top (\hat{\boldsymbol{\Omega}}_g^4 - \hat{\boldsymbol{\Omega}}_g^3) \mathbf{y}_{ig} \Big\}, \tag{S7}$$

and where $\bar{\lambda}_g = \frac{\lambda_{g\bullet}}{n_g}$. The derivation of this estimate is given in Section S3.3 below. One would then choose $\boldsymbol{\Lambda}^\star$ such that the FKL approximate cross-validation score is minimized:

$$\boldsymbol{\Lambda}^\star = \arg\min_{\boldsymbol{\Lambda}} \widehat{\text{FKL}}(\boldsymbol{\Lambda}), \quad \text{subject to:} \quad \boldsymbol{\Lambda} \geq \mathbf{0} \wedge \text{diag}(\boldsymbol{\Lambda}) > \mathbf{0}. \tag{S8}$$

The closed form expression in (S6) implies that $\boldsymbol{\Lambda}^\star$ is more rapidly determined than $\boldsymbol{\Lambda}^*$. As seen in the derivation, $\boldsymbol{\Lambda}^* \approx \boldsymbol{\Lambda}^\star$ for large sample sizes.

S3.3 Derivation

Here we give, borrowing some ideas from Vujačić et al. [42], the derivation of the estimate (S6). Let observation i in class g be denoted by $\mathbf{y}_{ig}$ and let $\mathbf{S} = \mathbf{S}_{ig} = \mathbf{y}_{ig}\mathbf{y}_{ig}^\top$ be the sample covariance or scatter matrix of that observation. As before, the singularly indexed $\mathbf{S}_g = \frac{1}{n_g}\sum_{i=1}^{n_g} \mathbf{S}_{ig}$ is the class-specific sample covariance matrix. Throughout this section we will conveniently drop (some of) the explicit notation.

The FKL divergence reframes the LOOCV score in terms of a likelihood evaluation and a bias term when $\mathbf{S}$ is *not* left out of class g. We thus study the change in the estimate as function of the single scatter matrix $\mathbf{S}$. Let $\hat{\boldsymbol{\Omega}}_g(\mathbf{S}) = \hat{\boldsymbol{\Omega}}_g^{\neg ig}$ be the estimate in class g when $\mathbf{S}$ is omitted. That is, $\hat{\boldsymbol{\Omega}}_g(\mathbf{S})$ is part of the solution to the system

$$\boldsymbol{\Omega}_a^{-1} + \mu_{aa}\boldsymbol{\Omega}_a + \mathbb{1}[a{=}g]\mathbf{S} + \sum_{b \neq a} \mu_{ab}\boldsymbol{\Omega}_b + \mathbf{A}_a = \mathbf{0}, \quad \text{for all} \quad a = 1, \ldots, G, \tag{S9}$$

where $\mu_{aa} = -\frac{\lambda_{a\bullet}}{n_a}$, $\mu_{ab} = \frac{\lambda_{ab}}{n_a}$, and where $\mathbf{A}_a$ is a matrix determined by the remaining data, penalty parameters and targets. Note that the penalized MLE can be denoted $\hat{\boldsymbol{\Omega}}_g = \hat{\boldsymbol{\Omega}}_g(\mathbf{0})$, which corresponds to the 'full' estimate resulting from the full gradient equation (28).

We wish to approximate $\hat{\boldsymbol{\Omega}}_g(\mathbf{S})$ by a Taylor expansion around $\hat{\boldsymbol{\Omega}}_g(\mathbf{0})$, i.e.:

$$\hat{\boldsymbol{\Omega}}_a(\mathbf{S}) \approx \hat{\boldsymbol{\Omega}}_a(\mathbf{0}) + \sum_{j,j'} \frac{\partial \hat{\boldsymbol{\Omega}}_a}{\partial S_{jj'}} S_{jj'}.$$

Differentiating (S9) w.r.t. $S_{jj'}$, the (j, j')th entry in $\mathbf{S}$, and equating to zero

yields

$$\begin{aligned}\mathbf{0} &= -\hat{\mathbf{\Omega}}_a^{-1}\frac{\partial\hat{\mathbf{\Omega}}_a}{\partial S_{jj'}}\hat{\mathbf{\Omega}}_a^{-1} + \mu_{aa}\frac{\partial\hat{\mathbf{\Omega}}_a}{\partial S_{jj'}} + \mathbb{1}[a{=}g]\mathbf{E}_{jj'} + \sum_{b\neq a}\mu_{ab}\frac{\partial\hat{\mathbf{\Omega}}_b}{\partial S_{jj'}} \\ &= -\hat{\mathbf{\Omega}}_a^{-1}\frac{\partial\hat{\mathbf{\Omega}}_a}{\partial S_{jj'}}\hat{\mathbf{\Omega}}_a^{-1} + \sum_{b}\mu_{ab}\frac{\partial\hat{\mathbf{\Omega}}_b}{\partial S_{jj'}} + \mathbb{1}[a{=}g]\mathbf{E}_{jj'}, \quad \text{for all} \quad j, j', \end{aligned} \tag{S10}$$

where $\mathbf{E}_{jj'}$ is the null matrix except for unity in entries (j, j') and (j', j). The third term is obtained as $\partial\mathbf{S}/\partial S_{jj'} = \mathbf{E}_{jj'}$ by the symmetric structure of $\mathbf{S}$. This is also seen from the fact that $\mathbf{S} = \sum_{jj'} S_{jj'}\mathbf{E}_{jj'}$. Let

$$\mathbf{V}(\mathbf{S})_a = \sum_{j,j'}\frac{\partial\hat{\mathbf{\Omega}}_a}{\partial S_{jj'}}S_{jj'},$$

and multiply (S10) by $S_{jj'}$ and sum over all j, j' to obtain

$$\hat{\mathbf{\Omega}}_a^{-1}\mathbf{V}(\mathbf{S})_a\hat{\mathbf{\Omega}}_a^{-1} - \sum_b \mu_{ab}\mathbf{V}(\mathbf{S})_b = \mathbb{1}[a{=}g]\mathbf{S}, \quad \text{for all} \quad a = 1, \ldots, G. \tag{S11}$$

We seek the solution vector $\mathbf{V} = \{\mathbf{V}(\mathbf{S})_a\}_{a=1}^G$ of square matrices for the system of equations in (S11) which can be rewritten in the following way. Introduce and consider the linear operator (or block matrix):

$$\mathbf{N} = \{\mathbf{N}_{ab}\}_{a,b=1}^G \quad \text{where} \quad \mathbf{N}_{ab} = \begin{cases}\hat{\mathbf{\Omega}}_a^{-1}\otimes\hat{\mathbf{\Omega}}_a^{-1} - \mu_{aa}\mathbf{I}_p\otimes\mathbf{I}_p & \text{if} \quad a = b \\ -\mu_{ab}\mathbf{I}_p\otimes\mathbf{I}_p & \text{if} \quad a \neq b\end{cases}.$$

Then $\mathbf{V}$ can be verified to be the solution to the system (S10) as

$$\begin{aligned}\mathbf{N}(\mathbf{V})_a &= \sum_b \mathbf{N}_{ab}\mathbf{V}(\mathbf{S})_b = \mathbf{0} \quad \text{for} \quad a \neq g, \quad \text{and} \\ \mathbf{N}(\mathbf{V})_g &= \sum_b \mathbf{N}_{gb}\mathbf{V}(\mathbf{S})_b = \mathbf{S} \quad \text{for} \quad a = g.\end{aligned}$$

Hence we need to invert $\mathbf{N}$ to solve for $\mathbf{V}$. The structure of $\mathbf{N}$ is relatively simple, but there seems to be no (if any) simple inverse. Note that $\mathbf{N} = \mathbf{D} - \mathbf{M}$ is the difference of a (block) diagonal matrix $\mathbf{D}$ and a matrix $\mathbf{M}$ depending on the μ's:

$$\begin{aligned}\mathbf{D}_{aa} &= \hat{\mathbf{\Omega}}_a^{-1}\otimes\hat{\mathbf{\Omega}}_a^{-1}, \\ \mathbf{M}_{ab} &= \mu_{ab}\mathbf{I}_p\otimes\mathbf{I}_p.\end{aligned}$$

In terms of the μ's we obtain to first order that

$$\mathbf{N}^{-1} = (\mathbf{D} - \mathbf{M})^{-1} \approx \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{M}\mathbf{D}^{-1},$$

yielding the approximation

$$\begin{aligned}\hat{\boldsymbol{\Omega}}_g(\mathbf{S}) &\approx \hat{\boldsymbol{\Omega}}_g + (\hat{\boldsymbol{\Omega}}_g \otimes \hat{\boldsymbol{\Omega}}_g + \mu_{gg}\hat{\boldsymbol{\Omega}}_g^2 \otimes \hat{\boldsymbol{\Omega}}_g^2)(\mathbf{S}) \\ &= \hat{\boldsymbol{\Omega}}_g + \hat{\boldsymbol{\Omega}}_g\mathbf{S}\hat{\boldsymbol{\Omega}}_g + \mu_{gg}\hat{\boldsymbol{\Omega}}_g^2\mathbf{S}\hat{\boldsymbol{\Omega}}_g^2, \end{aligned} \tag{S12}$$

where $\hat{\boldsymbol{\Omega}}_g = \hat{\boldsymbol{\Omega}}(\mathbf{0})$. To a first order in μ_{gg} this is the same as the approximation

$$\hat{\boldsymbol{\Omega}}_g(\mathbf{S}) \approx \hat{\boldsymbol{\Omega}}_g + (\hat{\boldsymbol{\Omega}}_g^{-1} \otimes \hat{\boldsymbol{\Omega}}_g^{-1} - \mu_{gg}\mathbf{I}_p \otimes \mathbf{I}_p)^{-1}(\mathbf{S}).$$

We also need an approximation for $\ln|\hat{\boldsymbol{\Omega}}_g(\mathbf{S})|$. By first-order Taylor expansion around $\mathbf{S} = \mathbf{0}$ we have

$$\begin{aligned}\ln|\hat{\boldsymbol{\Omega}}_g(\mathbf{S})| &\approx \ln|\hat{\boldsymbol{\Omega}}_g(\mathbf{0})| + \sum_{j,j'} \operatorname{tr}\Big[\hat{\boldsymbol{\Omega}}_g^{-1}(\mathbf{0})\frac{\partial\hat{\boldsymbol{\Omega}}_g}{\partial S_{jj'}}\Big]S_{jj'} \\ &\overset{(S12)}{\approx} \ln|\hat{\boldsymbol{\Omega}}_g(\mathbf{0})| + \operatorname{tr}\Big[\hat{\boldsymbol{\Omega}}_g^{-1}(\hat{\boldsymbol{\Omega}}_g \otimes \hat{\boldsymbol{\Omega}}_g + \mu_{gg}\hat{\boldsymbol{\Omega}}_g^2 \otimes \hat{\boldsymbol{\Omega}}_g^2)(\mathbf{S})\Big] \\ &= \ln|\hat{\boldsymbol{\Omega}}_g(\mathbf{0})| + \operatorname{tr}\big(\mathbf{S}\hat{\boldsymbol{\Omega}}_g + \mu_{gg}\hat{\boldsymbol{\Omega}}\mathbf{S}\hat{\boldsymbol{\Omega}}_g^2\big), \end{aligned} \tag{S13}$$

where we have used that $\frac{d}{dt}\ln|\mathbf{A}(t)| = \operatorname{tr}\big[\mathbf{A}(t)^{-1}\frac{d\mathbf{A}}{dt}\big]$ and $\frac{\partial\boldsymbol{\Omega}_g}{\partial S_{jj'}} \approx (\hat{\boldsymbol{\Omega}}_g \otimes \hat{\boldsymbol{\Omega}}_g + \mu_{gg}\hat{\boldsymbol{\Omega}}_g^2 \otimes \hat{\boldsymbol{\Omega}}_g^2)(\mathbf{E}_{jj'})$. We now have the necessary equations to derive the FKL approximate cross-validation score.

Define

$$f(\mathbf{A}, \mathbf{B}) = \ln|\mathbf{B}| - \operatorname{tr}(\mathbf{B}\mathbf{A}) \tag{S14}$$

by which the identity

$$\sum_{i=1}^{n_g} f(\mathbf{S}_{ig}, \boldsymbol{\Omega}_g) = n_g f(\mathbf{S}_g, \boldsymbol{\Omega}_g) \tag{S15}$$

holds for all g. The full likelihood (3) in terms of f is given by

$$\mathcal{L}(\{\boldsymbol{\Omega}_g\}; \{\mathbf{S}_g\}) \propto \sum_{g=1}^{G} \frac{n_g}{2}\Big\{\ln|\boldsymbol{\Omega}_g| - \operatorname{tr}(\boldsymbol{\Omega}_g\mathbf{S}_g)\Big\} = \sum_{g=1}^{G} \frac{n_g}{2} f(\mathbf{S}_g, \boldsymbol{\Omega}_g), \tag{S16}$$

while the likelihood of a single $\mathbf{S}_{ig}$ is

$$\mathcal{L}_{ig}(\boldsymbol{\Omega}_g; \mathbf{S}_{ig}) \propto \frac{1}{2}\Big\{\ln|\boldsymbol{\Omega}_g| - \operatorname{tr}(\boldsymbol{\Omega}_g\mathbf{S}_{ig})\Big\} = \frac{1}{2} f(\mathbf{S}_{ig}, \boldsymbol{\Omega}_g). \tag{S17}$$

In our setting, the fused LOOCV score is given by:

$$
\begin{aligned}
&\text{LOOCV} \\
&= -\frac{1}{n_\bullet}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\mathcal{L}_{ig}(\hat{\boldsymbol{\Omega}}_g^{\neg ig};\mathbf{S}_{ig}) \\
&\overset{(S17)}{=} -\frac{1}{n_\bullet}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\frac{1}{2}f(\mathbf{S}_{ig},\hat{\boldsymbol{\Omega}}_g^{\neg ig}) \\
&= -\frac{1}{n_\bullet}\sum_{g=1}^{G}\frac{1}{2}\sum_{i=1}^{n_g}\Big[f(\mathbf{S}_{ig},\hat{\boldsymbol{\Omega}}_g) + f(\mathbf{S}_{ig},\hat{\boldsymbol{\Omega}}_g^{\neg ig}) - f(\mathbf{S}_{ig},\hat{\boldsymbol{\Omega}}_g)\Big] \\
&\overset{(S15)}{=} -\frac{1}{n_\bullet}\sum_{g=1}^{G}\frac{n_g}{2}f(\mathbf{S}_g,\hat{\boldsymbol{\Omega}}_g) - \frac{1}{n_\bullet}\sum_{g=1}^{G}\frac{1}{2}\sum_{i=1}^{n_g}\Big[f(\mathbf{S}_{ig},\hat{\boldsymbol{\Omega}}_g^{\neg ig}) - f(\mathbf{S}_{ig},\hat{\boldsymbol{\Omega}}_g)\Big] \\
&\overset{(S16)}{=} -\frac{1}{n_\bullet}\mathcal{L}(\{\hat{\boldsymbol{\Omega}}_g\};\{\mathbf{S}_g\}) - \frac{1}{2n_\bullet}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\Big[f(\mathbf{S}_{ig},\hat{\boldsymbol{\Omega}}_g^{\neg ig}) - f(\mathbf{S}_{ig},\hat{\boldsymbol{\Omega}}_g)\Big] \\
&\overset{(S14)}{=} -\frac{1}{n_\bullet}\mathcal{L}(\{\hat{\boldsymbol{\Omega}}_g\};\{\mathbf{S}_g\}) - \frac{1}{2n_\bullet}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\Big[\ln|\hat{\boldsymbol{\Omega}}_g^{\neg ig}| - \operatorname{tr}(\hat{\boldsymbol{\Omega}}_g^{\neg ig}\mathbf{S}_{ig}) - \ln|\hat{\boldsymbol{\Omega}}_g| + \operatorname{tr}(\hat{\boldsymbol{\Omega}}_g\mathbf{S}_{ig})\Big].
\end{aligned}
$$

Now, substitution of (S12) and (S13) into the latter gives the FKL approximate cross-validation score as an approximation to the fused LOOCV score:

$$
\text{LOOCV} \approx \widehat{\text{FKL}} = -\frac{1}{n_\bullet}\mathcal{L}(\{\hat{\boldsymbol{\Omega}}_g\};\{\mathbf{S}_g\}) + \frac{1}{2n_\bullet}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\boldsymbol{\zeta}_{ig},
$$

where

$$
\begin{aligned}
\boldsymbol{\zeta}_{ig} &= \operatorname{tr}\big(\hat{\boldsymbol{\Omega}}\mathbf{S}\hat{\boldsymbol{\Omega}} + \mu_{gg}\hat{\boldsymbol{\Omega}}^2\mathbf{S}\hat{\boldsymbol{\Omega}}^2\big) - \operatorname{tr}\big(\mathbf{S}\hat{\boldsymbol{\Omega}} + \mu_{gg}\hat{\boldsymbol{\Omega}}\mathbf{S}\hat{\boldsymbol{\Omega}}^2\big) \\
&= \operatorname{tr}\big(\hat{\boldsymbol{\Omega}}\mathbf{S}\hat{\boldsymbol{\Omega}}\big) + \mu_{gg}\operatorname{tr}\big(\hat{\boldsymbol{\Omega}}^2\mathbf{S}\hat{\boldsymbol{\Omega}}^2\big) - \operatorname{tr}\big(\mathbf{S}\hat{\boldsymbol{\Omega}}\big) - \mu_{gg}\operatorname{tr}\big(\hat{\boldsymbol{\Omega}}\mathbf{S}\hat{\boldsymbol{\Omega}}^2\big) \\
&= \operatorname{tr}\big(\mathbf{S}\hat{\boldsymbol{\Omega}}^2\big) + \mu_{gg}\operatorname{tr}\big(\mathbf{S}\hat{\boldsymbol{\Omega}}^4\big) - \operatorname{tr}\big(\mathbf{S}\hat{\boldsymbol{\Omega}}\big) - \mu_{gg}\operatorname{tr}\big(\mathbf{S}\hat{\boldsymbol{\Omega}}^3\big) \\
&= \operatorname{tr}\big[\mathbf{S}(\hat{\boldsymbol{\Omega}}^2 - \hat{\boldsymbol{\Omega}})\big] + \mu_{gg}\operatorname{tr}\big[\mathbf{S}(\hat{\boldsymbol{\Omega}}^4 - \hat{\boldsymbol{\Omega}}^3)\big] \\
&= \mathbf{y}_{ig}^\top(\hat{\boldsymbol{\Omega}}^2 - \hat{\boldsymbol{\Omega}})\mathbf{y}_{ig} + \mu_{gg}\mathbf{y}_{ig}^\top(\hat{\boldsymbol{\Omega}}^4 - \hat{\boldsymbol{\Omega}}^3)\mathbf{y}_{ig}. \qquad \text{(S18)}
\end{aligned}
$$

To arrive at (S18) we have used the linear and cyclic properties of the trace operator. As $\mathbf{S} = \mathbf{y}_{ig}\mathbf{y}_{ig}^\top$, the cyclic property implies the final equality since $\operatorname{tr}(\mathbf{SA}) = \operatorname{tr}(\mathbf{y}_{ig}\mathbf{y}_{ig}^\top\mathbf{A}) = \operatorname{tr}(\mathbf{y}_{ig}^\top\mathbf{A}\mathbf{y}_{ig}) = \mathbf{y}_{ig}^\top\mathbf{A}\mathbf{y}_{ig}$. Equation (S18) is equivalent to the summand in (S7).

Paper V

Unaccounted Uncertainty from qPCR Efficiency Estimates Imply Uncontrolled False Positive Rates

Anders Ellern Bilgrau, Steffen Falgreen, Anders Petersen,
Malene Krag Kjeldsen, Julie Støve Bødker, Hans Erik Johnsen,
Karen Dybkær, and Martin Bøgsted

Submitted to
BMC Bioinformatics, 2015

Preface: This paper discusses unfortunate statistical methodology widely used to analyze so-called quantitative polymerase chain-reaction (qPCR) experiments.

The qPCR technology is used to quantify the abundance of particular nucleic acid of interest. In short, it works by repeatedly splitting and copying the DNA in the biological sample. This causes a chain reaction doubling the number of copies for each such so-called cycle. As florescent molecules are attached to the molecules of interest, this allows for quantifying the abundance in each cycle. The number of doublings (cycles) it takes to reach a certain florescence level is called the C_q-value which reflects the number of molecules in the original sample.

However, each cycle does not perfectly double the number of molecules. This biases the estimation of the C_q-value and derived quantities. To account for this imperfection, the *amplification efficiency* is estimated and used in the computations. The prevailing current statistical methods, however, do not account for the uncertainty of this amplification estimate and so the standard errors, p-values, and confidence intervals are potentially overly optimistic. This is particularly unfortunate since qPCR is an oft-used tool for validation. This paper serves to illustrate, discuss, and attempt to resolve this problem.

The manuscript, data, and R-scripts are available at

`https://github.com/AEBilgrau/effadj`

for easy reproduction of the formatted paper and its results.

 The layout has been revised.

Unaccounted Uncertainty from qPCR Efficiency Estimates Imply Uncontrolled False Positive Rates

Abstract

Background Accurate adjustment for amplification efficiency (AE) is an important part of real-time quantitative polymerase chain reaction (qPCR) experiments. The most commonly used correction strategy is to estimate the AE by dilution experiments and use this as a plug-in when efficiency correcting the $\Delta\Delta C_q$. Currently, it is recommended to determine the AE with high precision as this plug-in approach does not account for the AE uncertainty, implicitly assuming an infinitely precise AE estimate. Determining the AE with such precision, however, requires tedious laboratory work and vast amounts of biological material. Violation of the assumption leads to overly optimistic standard errors of the $\Delta\Delta C_q$, confidence intervals, and p-values which ultimately increase the type I error rate beyond the expected significance level. As qPCR is often used for validation it should be a high priority to account for the uncertainty of the AE estimate and thereby properly bounding the type I error rate and achieve the desired significance level.

Results We suggest and benchmark different methods to obtain the standard error of the efficiency adjusted $\Delta\Delta C_q$ using the statistical delta method, Monte Carlo integration, or bootstrapping. Our suggested methods are founded in a linear mixed effects model (LMM) framework, but the problem and ideas apply in all qPCR experiments. The methods and impact of the AE uncertainty are illustrated in three qPCR applications and a simulation study. In addition, we validate findings suggesting that *MGST1* is differentially expressed between high and low abundance culture initiating cells in multiple myeloma and that microRNA-127 is differentially expressed between testicular and nodal lymphomas.

Conclusions We conclude, that the commonly used efficiency corrected quantities disregard the uncertainty of the AE, which can drastically impact the standard error and lead to increased false positive rates. Our suggestions show that it is possible to easily perform statistical inference of $\Delta\Delta C_q$, whilst properly accounting for the AE uncertainty and better controlling the false positive rate.

1 Introduction

Despite being an aging technique, real-time quantitative polymerase chain reaction (qPCR)—arguably one of the most significant biotech discoveries of all time—is still heavily used in molecular biology [21]. qPCR is an extremely sensitive and cost-effective technique to amplify and quantitate the abundance of DNA and even mRNA by using so-called reverse transcriptase. In life sciences, qPCR is typically applied to quantify candidate gene transcripts that are biomarkers of diagnostic, prognostic, and even predictive value in e.g. infectious diseases and cancer. In the slip stream of high-volume omics-data,

another very important application of qPCR has arisen. Here qPCR is the gold standard validation tool for the most promising gene transcripts generated by high-throughput screening studies such as microarrays or sequencing. For validation experiments in particular the ability to control the type I error rate is very important. Unfortunately, important statistical details are often omitted resulting in a failure to obtain the desired type I error probability. Validation without such an ability cannot be considered very meaningful and therefore conservative approaches should be taken.

Like all experiments in molecular biology and chemistry, qPCR is sensitive not only to genes and gene transcripts of interest, but also laboratory settings and experimental conditions. There is an incredible amount of sources of systematic and non-systematic variation including temperature variations, concentration differences by pipetting errors, and primer affinity, in addition to the genuine biological variations of interest across case and control samples. Laboratory guidelines and increasingly sophisticated statistical modelling have been established to combat many of these systematic errors [5, 6, 28].

The so-called $\Delta\Delta C_q$ quantity is the normalized relative expression of a target gene of interest between treated (case) and untreated samples (control) accounting for undesired variations using one or more endogenous reference genes (also called housekeeping gene) assumed to be approximately unchanged due to the treatment. The $\Delta\Delta C_q$-value is usually based on the assumption of perfect AEs for both the target and reference gene. However, the target and reference genes might be subject to different AE which yield biased $\Delta\Delta C_q$-values. In turn, the $\Delta\Delta C_q$ has been modified to AE corrected versions [19, 22, 23].

Despite the tremendous success of qPCR, 'statistical inference considerations are still not accorded high enough priority' [5, 6]. We find this particular true for the estimation of the AE. Although efficiency calibration has been extensively treated by [19] or in the more generalized model by [34], there seems to be a lack of systematic studies of the unavoidable influence of the uncertainty of the AE estimate on the conclusions of qPCR experiments based on formal statistical inference. The current AE adjusted $\Delta\Delta C_q$ methods do not account for the *uncertainty of the estimated AE* and thus effectively assumes the AE to be estimated with infinite precision. This assumption implies a systematic underestimation of the standard error of $\Delta\Delta C_q$ leading to too narrow confidence intervals, decreased p-values, and thereby increased type I error rates. If the AE is poorly determined this underestimation can drastically increase the standard error of $\Delta\Delta C_q$ and similar quantities.

Nordgård et al. [18] studied error propagation including the effect of the AE uncertainty on the C_q-values. This study was, however, statistically informal and made no attempt to quantify the effect on the $\Delta\Delta C_q$ and inference hereof. Furthermore, Nordgård et al. [18] considered AE estimation from the amplification curve (thus for each sample) and not from separate dilution experiments. In this paper, we discuss only the AE from dilution curves explicitly. However, similar problems exist as the AE estimates from the amplification curves *also*

have an associated standard error which affect the 'down-steam' quantities and inference.

Svec et al. [31] recently assessed the impact of the AE uncertainty as a function of the number of technical replications at each concentration and the qPCR instrument. They conclude that a minimum of 3–4 replicates at each concentration are needed and that a significant inter qPCR instrument effect is present. However, they do not gauge the effect of the number of concentrations used—an important variable as additional technical replicates rarely contribute with much information to determine the AE. Nonetheless, Svec et al. [31] do not address the impact of the AE uncertainty on formal statistical inference on the $\Delta\Delta C_q$, as this paper intends.

1.1 Aims

Primarily, we aim to highlight the common problem of disregarding the uncertainty of the AE estimate in statistical inference of the $\Delta\Delta C_q$-value in qPCR experiments. And we propose and benchmark different off-the-shelf and novel solutions to this problem.

To this end, we employ a statistical model which allows such formal inference. This covers statistical model formulation, confidence intervals, hypothesis testing, and power calculation, with special emphasis on false positive rates. Simultaneous estimation of the uncertainty of the AE estimate and mean C_q-values by linear mixed effects models (LMM), which allows a more appropriate handling of the technical and sample errors, is described. We investigate the use of the statistical delta method, Monte Carlo integration, or bootstrapping to correctly perform inference on the value of $\Delta\Delta C_q$.

Note two important observations: First, the problem exists for *all* statistical models and methods which incorrectly disregard the uncertainty of the AE estimate and is not limited to LMMs. Second, the problem exists not only for $\Delta\Delta C_q$-values, but also all similar quantities, e.g. ΔC_q and C_q, and the statistical inferences based on these.

The idea of using LMMs for qPCR experiments is not new [1, 2, 9, 16, 29]. [2] and [1] have used mixed effects modeling to identify candidate normalizing genes. Fu et al. [9] applied the related generalized estimation equations to handle intra and inter group variation. However, the usage of LMMs combined with the statistical delta method, Monte Carlo integration, or bootstrapping to handle uncertainty stemming from the efficiency estimation seems to be novel and provides a general statistical framework for qPCR experiments and may be considered an extension of the strategy by Yuan et al. [34]. Steibel et al. [29] and Matz et al. [16] use the mixed models primarily for the C_q-value estimation.

We demonstrate that considering the uncertainty of the AE is, unsurprisingly, highly important when the AE is determined with inadequate precision and vice versa. We do so by three application examples and a simulation experiment. In the first two applications, the consideration of the AE uncertainty is largely unimportant for $\Delta\Delta C_q$ inference due to a large number of dilution steps

and well-determined AE. In the last application, we see that the AE uncertainties have a drastically different impact on $\Delta\Delta C_q$ inference. In a simulation study, we show that the methods proposed indeed control the false positive rate better than the conventional approach and provide further insight into the problem.

In the first application, we also verify that multiple myeloma cancer cell lines differentially express the *MGST1* gene depending on the abundance of culture initiating cells. In the second application, the approaches are also used to design and analyze a study which results turned out to support the hypothesis of [25] that miRNA-127 is differentially expressed between testicular and nodal DLBCL.

2 Methods

2.1 Observational model

In order to approximate the standard error of the AE adjusted $\Delta\Delta C_q$ we model the amplification process in the following way

$$F_{C_q} = \kappa\sigma N_0 (2^{\alpha})^{C_q}, \tag{1}$$

where F_{C_q} is the fluorescence measurement at the C_q'th cycle, κ is a proportionality constant, N_0 is the number of transcripts of interest in the initial sample before amplification, σ is the sample specific handling error and α is the percentage of the $\log_2$-AE. In practice, the transcript abundance level is determined by the cycle C_q for which a given fluorescence measurement F_{C_q} is reached. We rearrange (1) and notice that C_q can be expressed as $\alpha C_q = \log_2 F_{C_q} - \log_2 \kappa\sigma N_0$. In order to estimate the relative amount of target (tgt) gene transcripts between case and control (ctrl) samples, we assume the amount of the reference (ref) gene template is the same in both the case and the control, $N_{0,\text{ref,case}} = N_{0,\text{ref,ctrl}}$, and that the AE only vary between the target and reference gene. We then arrive at the following expression for the $\log_2$-fold change of the target gene template between case and controls:

$$\begin{aligned}
\log_2\Big(\frac{N_{0,\text{tgt,case}}}{N_{0,\text{tgt,ctrl}}}\Big) &= \log_2 \kappa\sigma_{\text{case}} N_{0,\text{tgt,case}} - \log_2 \kappa\sigma_{\text{case}} N_{0,\text{ref,case}} \\
&\quad - \log_2 \kappa\sigma_{\text{ctrl}} N_{0,\text{tgt,ctrl}} + \log_2 \kappa\sigma_{\text{ctrl}} N_{0,\text{ref,ctrl}} \\
&= -\big\{(\alpha_{\text{tgt}} C_{q,\text{tgt,case}} - \alpha_{\text{ref}} C_{q,\text{ref,case}}) \\
&\quad - (\alpha_{\text{tgt}} C_{q,\text{tgt,ctrl}} - \alpha_{\text{ref}} C_{q,\text{ref,ctrl}})\big\},
\end{aligned}$$

assuming that the C_q-values have been determined by a common florescence level F_{C_q}. This method of estimating the log relative abundance between case and controls is often referred to as the $\Delta\Delta C_q$-method [15], after the double difference appearing in the expression:

$$\Delta\Delta C_q := (\alpha_{\text{tgt}} C_{q,\text{tgt,case}} - \alpha_{\text{ref}} C_{q,\text{ref,case}}) - (\alpha_{\text{tgt}} C_{q,\text{tgt,ctrl}} - \alpha_{\text{ref}} C_{q,\text{ref,ctrl}}). \tag{2}$$

Thus we have $2^{-\Delta\Delta C_q}$ as the relative abundance of the original target transcript corrected for the AE.

2.2 Statistical model

We study the problematic aspects of ignoring the uncertainty of the AE estimate. Note, however, that this problem persists for *all* statistical models and methods which naïvely 'plug-in' the AE estimate from dilution curves into formulae concerning the $\Delta\Delta C_q$.

For ease of notation we use the abbreviations $i \in \{\text{tgt}, \text{ref}\}$ for gene types target and reference; $j \in \{\text{case}, \text{ctrl}, \text{std}\}$ for sample types case, control, and standard curve; $s \in \{1, \ldots, n_{ij}\}$ for samples in the ij'th group; and $k \in \{0, \ldots, K_{ijs}\}$ for dilution steps for each sample. To estimate $\Delta\Delta C_q$ of (2), estimates of α_i are needed. A popular way of estimating the AE is by ordinary linear regression. I.e. by regressing $C_{q,ij}$ against a series of increasing values $0 = x_1 < \cdots < x_K$, defined by $N_{0,ijk} = N_{0,ij} 2^{-x_k}$, and naïvely plugging $\hat{\alpha}_i$ into (2) and thus disregarding its uncertainty. Here, k denotes the dilution step and x_k the number of 2-fold dilutions (e.g. $x_1 = 1$ means the first dilution step halves the original concentration). The estimation of the expected $C_{q,ij}$-values and α_i can then be estimated simultaneously when formulated as a LMM [20];

$$C_{q,ijsk} = \mu_{ij} + A_{js} + \gamma_i x_k + \epsilon_{ijsk}, \tag{3}$$

where A_{js} is a random effect from sample s under the j'th sample type, $\gamma_i = \alpha_i^{-1}$, and μ_{ij} is the group means. The random effects A_{js} are $\mathcal{N}(0, \sigma_S^2)$-distributed and the error terms ϵ_{ijsk} are independent and $\mathcal{N}(0, \sigma_j^2)$-distributed with a sample type specific variance σ_j^2. The random effects account for the paired samples across tgt/ref for each j. LMMs provide a more correct quantification of the sources of variation and thereby a more correct estimate of the uncertainty of μ_{ij} and their derived quantities.

In one application we shall relax the assumption that the AE is independent of j and consider group-specific AEs $\alpha_{ij} = \gamma_{ij}^{-1}$.

Although, variation due to technical replicates should be modeled in (3) as an additional random effect term, we average out technical replicates for simplicity. For further simplicity of this paper, we refrained from using multiple reference genes simultaneously in the $\Delta\Delta C_q$ estimation although our the framework and methods easily extends to this case.

2.3 Inference for $\Delta\Delta C_q$ by the delta method and Monte Carlo integration

We first consider hypothesis testing and confidence intervals for $\Delta\Delta C_q$ by the statistical delta method. Let the maximum likelihood estimates of the fixed effects

$$\boldsymbol{\theta} = (\mu_{\text{tgt,case}}, \mu_{\text{tgt,ctrl}}, \gamma_{\text{tgt}}, \mu_{\text{ref,case}}, \mu_{\text{ref,ctrl}}, \gamma_{\text{ref}})^\top$$

be denoted by $\hat{\boldsymbol{\theta}} = (\hat{\mu}_{\text{tgt,case}}, \hat{\mu}_{\text{tgt,ctrl}}, \hat{\gamma}_{\text{tgt}}, \hat{\mu}_{\text{ref,case}}, \hat{\mu}_{\text{ref,ctrl}}, \hat{\gamma}_{\text{ref}})^\top$. We wish to test the hypothesis $H_0 : c(\boldsymbol{\theta}) = 0$, where c is the continuously differentiable function of the fixed effects given by

$$c(\boldsymbol{\theta}) = \big\{(\mu_{\text{tgt,case}}\gamma_{\text{tgt}}^{-1} - \mu_{\text{ref,case}}\gamma_{\text{ref}}^{-1}) - (\mu_{\text{tgt,ctrl}}\gamma_{\text{tgt}}^{-1} - \mu_{\text{ref,ctrl}}\gamma_{\text{ref}}^{-1})\big\}. \tag{4}$$

The main task of this paper is to approximate the standard error of $c(\hat{\boldsymbol{\theta}})$ and thereby account for the uncertainty of $\Delta\Delta C_q$. That is, the standard error,

$$\text{se}(\hat{\boldsymbol{\theta}}) = \sqrt{\mathbb{V}\text{ar}\big[c(\hat{\boldsymbol{\theta}})\big]}, \tag{5}$$

is of central interest. The standard error is used in the statistic for testing H_0 given by $t = c(\hat{\boldsymbol{\theta}})/\text{se}(\hat{\boldsymbol{\theta}})$, which according to a first order Taylor series expansion of c can be approximated by

$$t = \frac{c(\hat{\boldsymbol{\theta}})}{\sqrt{\nabla c(\hat{\boldsymbol{\theta}})^\top \mathbb{V}\text{ar}[\hat{\boldsymbol{\theta}}] \nabla c(\hat{\boldsymbol{\theta}})}}. \tag{6}$$

According to Pinheiro and Bates [20, Section 2.4.2], t is approximately t-distributed with η degrees of freedom. The degrees of freedom of multilevel mixed effects models are non-trivial to obtain in general. We do not pursue this further and restrict ourselves to the case of balanced experimental designs where η is obtained relatively straight-forwardly.

On the basis of (6), an approximate $(1-\alpha)100\%$ confidence interval of $c(\boldsymbol{\theta})$ can then be given by

$$c(\hat{\boldsymbol{\theta}}) \pm t_{\alpha/2,\eta} \sqrt{\nabla c(\hat{\boldsymbol{\theta}})^\top \mathbb{V}\text{ar}[\hat{\boldsymbol{\theta}}] \nabla c(\hat{\boldsymbol{\theta}})}.$$

Likewise, p-values can be obtained by computing $P(|t| > T)$ where T is t-distributed with η degrees of freedom.

Alternatively to (6), the variance $\mathbb{V}\text{ar}\big[c(\hat{\boldsymbol{\theta}})\big]$ can be evaluated by Monte Carlo integration. One way is to simulate a large number N of parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ from a multivariate normal distribution using the estimated parameters $\mathcal{N}_6(\hat{\boldsymbol{\theta}}, \mathbb{V}\text{ar}[\hat{\boldsymbol{\theta}}])$ and compute the sample variance of $c(\boldsymbol{\theta}_1), \ldots, c(\boldsymbol{\theta}_N)$.

Both maximum likelihood (ML) and restricted maximum likelihood estimation (REML) of LMMs is implemented in the R-packages `lme4` and `nlme` [3, 20]. The packages readily provides the estimate $\hat{\boldsymbol{\theta}}$ and $\mathbb{V}\text{ar}[\hat{\boldsymbol{\theta}}]$ and we use these in the construction of test and confidence intervals for the $\Delta\Delta C_q$ as described above. The needed gradient in (6) is computed straight-forwardly from (4).

We note that the division by γ_j in (4) is problematic as $\hat{\gamma}_j$ is normally distributed and values near zero can increase the variance dramatic. In practice, this is only problematic if the standard error of $\hat{\gamma}_j$ is sufficiently large. One way to solve this problem is to use the $\log_2$ concentration as the response and the C_q-values as the explanatory variables in a regression model of the standard curve to estimate α_j directly. This approach is not without conceptual problems

as this puts the errors on the explanatory variables. To this end, note that the hypothesis $H_0 : \gamma_{\text{tgt}}\gamma_{\text{ref}}c(\hat{\boldsymbol{\theta}}) = 0$, can be equivalently tested for which the standard error of the test-statistic can be worked out exactly.

If γ_{tgt}^{-1} and γ_{ref}^{-1} are assumed to be one (or otherwise known) then (4) becomes a simple linear hypothesis for which the standard error is easily calculated. This corresponds to leaving out the terms in (3) involving these parameters and thus ignoring dilution data. If $\gamma_{\text{tgt}}^{-1} = \gamma_{\text{ref}}^{-1} = 1$ is assumed, we shall refer to the obtained estimate as the naïve LMM. If γ_{tgt}^{-1} and γ_{ref}^{-1} are assumed known (i.e. disregarding the standard error hereof) we refer to the obtained estimate as the efficiency corrected (EC) estimate. The estimate where the uncertainty of the AE is considered is referred to as efficiency corrected and variance adjusted by either the delta method (EC&VA1) or Monte Carlo integration (EC&VA2).

2.4 Inference for $\Delta\Delta C_q$ by the bootstrap method

We now consider hypothesis testing and confidence intervals for $\Delta\Delta C_q$ by bootstrapping as an alternative approach. The bootstrap, which avoids calculating gradients, is often cited to perform better in small sample situations [8].

The basic idea of the bootstrap is that inference on $\Delta\Delta C_q$ can be conducted by re-sampling the sample data and performing inference on the re-sampled data. In the usual qPCR setup with paired samples and dilution data, straightforward bootstrapping will quickly fail. We propose non-parametric block bootstrap samples for the case-control data generated by sampling matched pairs of tgt/ref genes with replacement for cases and controls, respectively. However, as we have only got a single observation for each dilution step we chose to re-sample residuals from a simple linear regression model and subsequently adding the residuals to the fitted values from the linear regression. Hence the B bootstrapped datasets consists of the re-sampled matched pairs and the residual bootstrapped standard curve. For each dataset, $\hat{\delta}_1 = \Delta\Delta C_q^{(1)}, \ldots, \hat{\delta}_B = \Delta\Delta C_q^{(B)}$ are computed to obtain the bootstrap distribution from which confidence intervals and p-values can be obtained. The standard error of $\Delta\Delta C_q$ is estimated by the sample standard deviation of the bootstrap distribution. A $(1-\alpha)100\%$ confidence interval can be computed as $(\hat{\delta}_{(\alpha/2)}, \hat{\delta}_{(1-\alpha/2)})$ where e.g. $\hat{\delta}_{(\alpha/2)}$ denotes the $\alpha/2$-percentile of $\hat{\delta}_1, \ldots, \hat{\delta}_B$. The p-value for the null hypothesis of $\delta = 0$ is computed by

$$2\min(\pi, 1-\pi) \text{ where } \pi = \frac{1 + \sum_{i=1}^{B} \mathbb{1}[\hat{\delta}_i \leq 0]}{B+1}.$$

While the bootstrap is an intuitive and excellent method for estimating the standard error, it quickly becomes computationally heavy. The rather complicated designs of qPCR experiments with paired samples, dilution data, and other random effects also makes the bootstrap less attractive as good bootstrap sampling schemes are hard to produce.

Alternatively, parametric bootstrap can be used by simulating datasets from

the fitted model. Here, both new random effects and noise terms are realized and added to the fitted values to generate new datasets.

3 Applications

We applied the described approaches to two qPCR validation experiments regarding culture initiating cells (CICs) in multiple myeloma (MM) and non-coding microRNAs in diffuse large B-cell lymphoma (DLBCL). In both experiments, the C_q-values were extracted for both the reference and target transcripts with automatic baseline and threshold selection [17]. We also illustrate the method on a public available qPCR dataset concerning the differential gene expression in arabidopsis thaliana grown under different conditions. In order to gauge the performance of the methods we subsequently performed a simulation study.

3.1 CIC study

Introduction

A cell is culture initiating if it can initiate a sustained production of cells when cultured in vitro. The viability potential of a cell population can be assessed by measuring the number of culture initiating cells (CICs). This number can be estimated by a dilution experiment where cells are seeded in decreasing numbers. The ratio of CICs can then be estimated by e.g. Poisson regression [13]. CICs are of particular interest in cancer research as cancers with high culture initiating potential seemingly have stem cell like properties making them resistant towards chemotherapy [7].

In search for genes associated with a high culture initiating potential in MM we made limiting dilution experiments of 14 MM cell lines and divided them into 7 cell lines with low and 7 cell lines with high culture initiating potential. Gene expression profiling by microarrays identified genes *MGST1* and *MMSET* to be differentially expressed between cell lines with high and low abundance of CICs. As gene expression detection by microarrays can be hampered by high false positive rates, the purpose of this experiment was to validate the findings of the association of *MGST1* and *MMSET* with culture initiating potential by qPCR.

Sample and data preparation

For this, 8 MM cell lines (AMO-1, KMM-1, KMS-11, KMS-12-PE, KMS-12-BM, MOLP-8, L-363, RPMI-8226) with $> 10\%$ CICs, and 8 MM cell lines (ANBL-1, KAS-6-1, LP-1, MOLP-2, NCI-H929, OPM-2, SK-MM-2, U-266) with $< 1\%$ CICs were used. The fraction of CICs was determined by the limiting dilution method, see [13]. Total RNA was isolated from frozen cell culture pellets, using a combined method of Trizol (Invitrogen) and Mirvana

spin columns (Ambion). Isolated RNA was reversed transcribed into complementary DNA (cDNA) synthesis using SuperScript III First-Strand Synthesis Supermix (Invitrogen). As input into the total cDNA synthesis of 250ng total RNA was used. Equal amounts of random hexamers and oligo(dT) were used as primers. Quantitative real-time reverse transcriptase polymerase chain reaction was performed on a Mx3000p qPCR system (Agilent Technologies/Stratgene) using the TaqMan UniversalPCR Master Mix, No AmpErase UNG, and TaqMan gene expression Assays (Applied Biosystems). The following TaqMan Gene Expression Assays were used (Assay ID numbers in parentheses): *MGST1* (Hs00220393_m1), *MMSET* (Hs00983716_m1). The two reference genes beta-actin (*ACTB*) and *GAPDH* were used as endogenous controls, assay IDs 4333762-0912030 and 4333764-1111036, respectively. For each target and reference transcripts a standard curve based on seven 2-fold dilutions was constructed on a reference sample consisting of material from the AMO-1 cell line.

3.2 DLBCL study

Introduction

The association between oncogenesis and micro RNAs (miRNAs), short non-coding RNA transcripts with regulatory capabilities, has recently prompted an immense research activity. The possibility to change treatment strategies by transfecting antisense oligonucleotide to control abnormally up-regulated miRNAs in malignant tissue is of particular interest [10]. In that respect up-regulated *miR-127* and *miR-143* in testicular DLBCL have shown treatment changing potential [25]. However, as the number of screened miRNAs was high and the sample size was low in Robertus et al.'s work invoking high risk of false discoveries we set out to validate the differential expression of *miR-127* and *miR-143* in tissues from our own laboratory using our improved qPCR analysis workflow.

Sample and data preparation

For this study, DLBCL samples were collected from 8 testicular (case) and 8 nodal (control) paraffin embedded lymphomas at Aalborg University Hospital. The samples were collected in accordance with a research protocol accepted by the Health Research Ethics Committee for North Denmark Region (No. N-20100059). Total RNA was isolated using a combined method of Trizol (Invitrogen) and Mirvana spin columns (Ambion). An amount of 10ng total RNA was synthesized into first strand cDNA in a 15μL reaction using TaqMan MicroRNA Reverse Transcription Kit (Applied Biosystems) according to the manufactures instruction. In total 1.33μL cDNA was used as template in the real time PCR amplification performed by Mx3000p QPCR system (Agilent Technologies/Stratgene) with sequence specific TaqMan primers (Applied Biosystems). As reference transcripts we chose *RNU-6B* and *RNU-24*, which

were less variable and equally expressed across nodal and extra-nodal samples among a larger list of candidate reference genes. For each target and reference transcripts a standard curve based on seven 2-fold dilutions was constructed on a reference sample consisting of pooled material from all 16 lymphomas.

3.3 Arabidopsis thaliana study

Introduction

In order to illustrate the effect of applying variance approximations in a dataset with a limited number of dilution steps and samples we considered the arabidopsis thaliana dataset published by Yuan et al. [34]. The dataset contains one gene of interest, *MT7*, potentially differentially expressed under two growth conditions of the plant arabidopsis thaliana and two reference genes ubiquitin (*UBQ*) and tublin.

Sample and data preparation

The arabidopsis thaliana plant growth, RNA extraction, and qPCR experiments were carried out as described in Yang et al. [33]. The cDNA was diluted into 1-to-4 and 1-to-16 serial dilutions. Real-time PCR experiments was performed in duplicates for each concentration [34].

Due to the the study design, we naturally fitted estimation efficiencies $\gamma_{ij} = \alpha_{ij}^{-1}$ for each group. Because of the few samples we omitted the, in this case, meaningless random sample effect of the LMM.

3.4 Simulation study

In order to properly benchmark statistical test procedures one needs to have an idea of the false positive rate (FPR), or type I error rate, as well as the true positive rate (TPR), or sensitivity. As ground truth is usually not available in non-synthetic data, we use simulation experiments to determine the error rates of the discussed statistical procedures.

In our setting, the FPR of a statistical test is the probability that the test incorrectly will declare a result statistically significant given a vanishing effect size or difference of $c(\theta) = 0$ between case and controls; i.e. $\text{FPR} = P\big(|t| > t_{1-\alpha/2,\eta} \;\big|\; c(\theta) = 0\big)$. On the other hand the TPR of the statistical test is the probability that the test will correctly declare a result statistically significant given an non-zero effect size $\delta = c(\theta)$ between case and controls; i.e. $\text{TPR} = P\big(|t| > t_{1-\alpha/2,\eta} \;\big|\; c(\theta) = \delta\big)$.

A straightforward way to obtain an estimate of the TPR is to simulate a large number n of datasets under the alternative hypothesis of $c(\theta) = \delta$, fit the model for each dataset, and compute t-values $t_1, \ldots, t_n$. From these t-scores

the TPR can estimated by

$$\widehat{\mathrm{TPR}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\big[|t_i| > t_{1-\alpha/2,\eta}\big],$$

where $\mathbb{1}[\,\cdot\,]$ is the indicator function. Hence, the estimated TPR is the fraction of tests correctly declared significant.

Likewise, an estimate of the FPR is obtained by simulating n datasets under the null hypothesis of $c(\theta) = 0$ and obtaining t-values $t_1, \ldots, t_n$ from which FPR is estimated by

$$\widehat{\mathrm{FPR}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\big[|t_i| > t_{1-\alpha/2,\eta}\big],$$

i.e. the fraction of tests incorrectly declared significant.

Based on the above statistical model, we estimate the FPR and the TPR for each discussed method under different choices of sample sizes and number of dilutions whilst fixing the size of the sample and experimental variations.

4 Results

4.1 CIC study

The C_q-values and dilution curves for the CIC study are depicted in Fig. 1 panels A–B, respectively. The simple linear regressions show well-determined standard curves with small standard errors on the estimate of the slopes.

The values of the considered estimators for $\Delta\Delta C_q$ are seen in Table 1. The table also shows results of tests for difference in gene expression assessed by the $\Delta\Delta C_q$ for both target genes *MGST1* and *MMSET* normalized to each of the reference genes *GAPDH* and *ACTB*. We used four different methods to estimate and perform inference: (1) EC: Efficiency corrected LMM estimate ignoring the uncertainty of the efficiency estimates. (2) EC&VA1: EC and variance adjusted LMM estimate using 1. order approximation. (3) EC&VA2: EC and variance adjusted LMM estimate using Monte Carlo integration. (4) Bootstrap: Estimate by the bootstrap described in Section 2.4 fitting the LMM and using the EC estimate.

Consider the first section of Table 1 where tgt *MGST1* is normalized against the reference *GAPDH*. The tests for a vanishing $\Delta\Delta C_q$ are all highly significant with comparable 95% CIs. As expected, the efficiency corrected estimates are unchanged due to the variance adjustment, and only the standard deviation of the estimate is increased. The increase of the standard error is very small resulting in small but unimportant increases of the absolute t- and p-values. The results remain significant for the *MGST1* gene. Very similar results are obtained if *ACTB* is used as reference. In conclusion, there is good evidence that *MGST1* is differentially expressed between cell lines with high and low abundance of CICs.

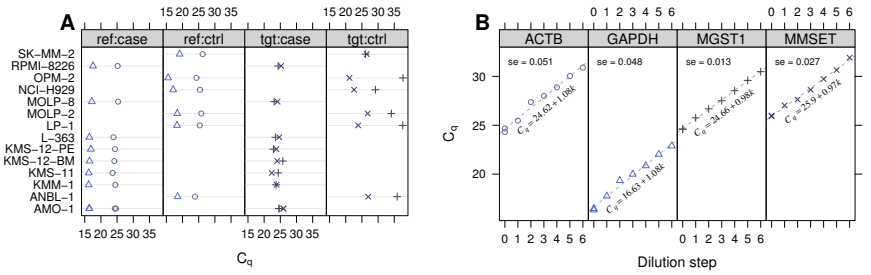


Fig. 1: Overview of CIC experiment data. A: Raw C_q-values for different cell lines (samples) for each gene type and sample type. The point type and colour differentiates the different gene types. B: Dilution data for reference genes (*ACTB*, *GAPDH*) and target genes (*MGST1*, *MMSET*).

For the target gene *MMSET* normalized with respect to both reference genes, all estimates are not significantly different from zero. Again, the various methods all agree and no substantial inter-method differences are seen and we find no evidence for differential expression of *MMSET* between cell lines with high and low abundance of CICs.

In all instances, our bootstrap scheme provides a standard deviation larger than what is obtained using the delta or Monte Carlo methods. The mean of the bootstrap distribution seems consistently larger that the other methods.

We see that the large number of dilution steps, as recommended and expected, ensures a low impact of the AE on the standard error and thus on the inference of the $\Delta\Delta C_q$.

4.2 DLBCL study

The C_q-values and dilution curves for the DLBCL study are depicted in Fig. 2, panels A–B, respectively. Analogous to the previous section, the differences in gene expressions assessed by the $\Delta\Delta C_q$ for the target genes *miR-127* and *miR-143* with respect to each reference gene rnu6b and rnu24 were estimated using the four different methods. Again 2000 bootstrapped samples were used. The results are seen in Table 2.

We notice the efficiency corrected estimates are exactly equal with and without variance adjustment, while the standard deviation of the estimate and the p-values are higher for the adjusted values as expected. The size of the increase is again undramatic hinting at well determined AE using the dilution curves.

For all combinations of reference genes the estimates for *miR-127* are significantly different from zero at the usual 5 % significant level, but not at the 1 % significance level. The *miR-143* estimates are not significantly different from zero. Despite the very small increase in standard error, the p-values increase

Table 1: CIC data: Method comparison for estimating the $\Delta\Delta C_q$-value (Est.) and the standard error (SE). EC denotes use of the plugin-estimator. VA denotes that the efficiency correction was variance adjusted using the delta method (1) or Monte Carlo integration (2). Bootstrap shows the mean and standard deviation of 2000 bootstrap samples using the EC estimate. The last two columns show the 95% lower and upper confidence interval limits.

	Est.	SE	t-value	df	p-value	LCL	UCL
MGST1 vs GAPDH							
EC	-8.622	1.625	-5.307	21	$2.915 \cdot 10^{-5}$	-12	-5.243
EC&VA1	-8.622	1.663	-5.184	21	$3.886 \cdot 10^{-5}$	-12.08	-5.163
EC&VA2	-8.622	1.669	-5.168	21	$4.040 \cdot 10^{-5}$	-12.09	-5.152
Bootstrap	-8.659	2.059			$9.995 \cdot 10^{-4}$	-12.49	-4.405
MGST1 vs ACTB							
EC	-8.984	1.612	-5.572	21	$1.574 \cdot 10^{-5}$	-12.34	-5.631
EC&VA1	-8.984	1.648	-5.452	21	$2.077 \cdot 10^{-5}$	-12.41	-5.557
EC&VA2	-8.984	1.649	-5.447	21	$2.104 \cdot 10^{-5}$	-12.41	-5.554
Bootstrap	-8.98	2.093			$9.995 \cdot 10^{-4}$	-12.75	-4.478
MMSET vs GAPDH							
EC	0.6793	0.5852	1.161	21	$2.588 \cdot 10^{-1}$	-0.5378	1.896
EC&VA1	0.6793	0.587	1.157	21	$2.601 \cdot 10^{-1}$	-0.5413	1.9
EC&VA2	0.6793	0.5888	1.154	21	$2.616 \cdot 10^{-1}$	-0.5452	1.904
Bootstrap	0.6878	0.6778			$3.118 \cdot 10^{-1}$	-0.6563	1.999
MMSET vs ACTB							
EC	0.318	0.9616	0.3308	21	$7.441 \cdot 10^{-1}$	-1.682	2.318
EC&VA1	0.318	0.9621	0.3306	21	$7.442 \cdot 10^{-1}$	-1.683	2.319
EC&VA2	0.318	0.9645	0.3298	21	$7.448 \cdot 10^{-1}$	-1.688	2.324
Bootstrap	0.3423	0.9872			$7.046 \cdot 10^{-1}$	-1.676	2.135

at the second digit.

The bootstrap method provides a standard deviation similar to the delta method and Monte Carlo integration for both *miR-127* and *miR-143*.

Regarding the biological interest, we conclude there is evidence for a difference in *miR-127* expression between testicular and nodal DLBCL whilst the data is not compatible with difference in *miR-143* expression. While the AE estimate had no influence in these cases a change in significance is easily imagined in other cases.

4.3 Arabidopsis thaliana data

The C_q-values and dilution data for the arabidopsis thaliana data are shown in Fig. 3.

The estimated difference in gene expression between case and control of the target gene *MT7* normalized to either reference (Tublin or *UBQ*) is seen in Table 3. The table shows the efficiency corrected method with and without variance adjustment by the delta method. In both cases, we see a dramatic increase in the standard error, p-values, and size of the confidence intervals. When using variance adjustment there is no longer a highly statistical significant difference in *MT7* expression between case and ctrl growth conditions.

The results may be surprising at first sight when considering the relatively

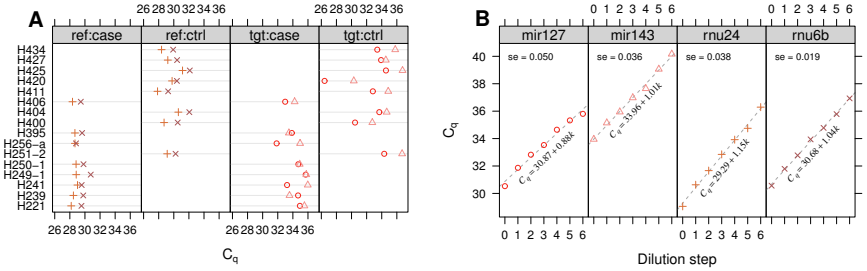


Fig. 2: Overview of DLBCL testis experiment data. A: Raw C_q-values for different patient samples for each gene type and sample type. The point type and colour differentiates the different gene types. B: Dilution data for reference genes (*RNU-24*, *RNU-6B*) and target genes (*miR-127*, *miR-143*).

Table 2: Testis data: Method comparison for estimating the $\Delta\Delta C_q$-value (Est.) and the standard error (SE). EC denotes use of the plugin-estimator. VA denotes that the efficiency correction was variance adjusted using the delta method (1) or Monte Carlo integration (2). Bootstrap shows the mean and standard deviation of 4 bootstrap samples using EC estimate. The last two columns show the 95% lower and upper confidence interval limits.

	Est.	SE	t-value	df	p-value	LCL	UCL
mir127 vs rnu6b							
EC	2.671	1.126	2.372	22	$2.684 \cdot 10^{-2}$	0.3361	5.005
EC&VA1	2.671	1.128	2.368	22	$2.711 \cdot 10^{-2}$	0.3315	5.01
EC&VA2	2.671	1.131	2.362	22	$2.747 \cdot 10^{-2}$	0.3253	5.016
Bootstrap	2.681	1.047			$9.995 \cdot 10^{-4}$	0.8756	4.817
mir127 vs rnu24							
EC	2.384	1.084	2.199	22	$3.868 \cdot 10^{-2}$	0.1357	4.631
EC&VA1	2.384	1.087	2.193	22	$3.915 \cdot 10^{-2}$	0.1298	4.637
EC&VA2	2.384	1.088	2.19	22	$3.941 \cdot 10^{-2}$	0.1265	4.641
Bootstrap	2.423	1.177			$9.995 \cdot 10^{-3}$	0.4164	5.022
mir143 vs rnu6b							
EC	1.165	0.846	1.377	22	$1.823 \cdot 10^{-1}$	-0.5893	2.92
EC&VA1	1.165	0.8463	1.377	22	$1.824 \cdot 10^{-1}$	-0.5899	2.92
EC&VA2	1.165	0.8475	1.375	22	$1.830 \cdot 10^{-1}$	-0.5923	2.923
Bootstrap	1.151	0.7943			$1.439 \cdot 10^{-1}$	-0.3409	2.7
mir143 vs rnu24							
EC	0.8781	0.8099	1.084	22	$2.900 \cdot 10^{-1}$	-0.8015	2.558
EC&VA1	0.8781	0.8101	1.084	22	$2.901 \cdot 10^{-1}$	-0.8019	2.558
EC&VA2	0.8781	0.8107	1.083	22	$2.905 \cdot 10^{-1}$	-0.8032	2.559
Bootstrap	0.8966	0.8216			$2.669 \cdot 10^{-1}$	-0.603	2.58

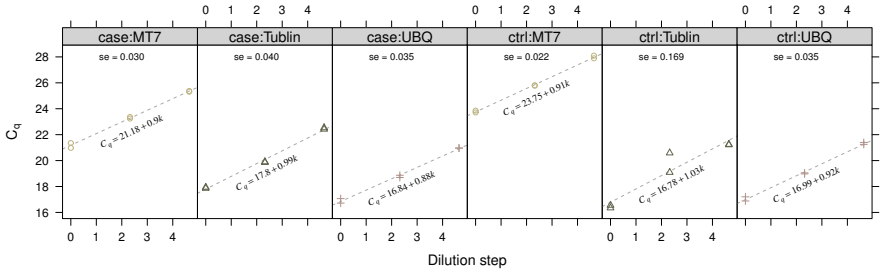


Fig. 3: Overview of Yuan et al. [34] experiment data. C_q-values against the dilution step for case and control samples. Dilution data are present for both the target ($MT7$) and reference genes (Tublin, UBQ).

Table 3: Yuan et al. [34] data: Method comparison for estimating the $\Delta\Delta C_q$-value (Est.) and the standard error (SE). EC denotes use of the plugin-estimator. VA denotes that the efficiency correction was variance adjusted using the delta method (1).

	Est.	SE	t-value	df	p-value	LCL	UCL
MT7 vs Tublin							
EC	-4.374	0.748	-5.847	4	$4.268 \cdot 10^{-3}$	-6.45	-2.297
EC&VA1	-4.374	6.561	-0.6666	4	$5.415 \cdot 10^{-1}$	-22.59	13.84
MT7 vs UBQ							
EC	-3.381	0.1682	-20.1	4	$3.616 \cdot 10^{-5}$	-3.848	-2.914
EC&VA1	-3.381	1.658	-2.039	4	$1.111 \cdot 10^{-1}$	-7.985	1.224

small standard errors of the slopes in the simple linear regressions shown in Fig. 3. One might imagine that the uncertainty of the AE is negligible and thus perform the usual analysis. However, we see the contrary for several reasons. First, using only 3 dilutions steps leaves very few degrees of freedom left in each group as we are left with few samples and a high number of parameters to be estimated. Secondly, as dilution curves are used for each group the four group-specific AE estimates will all contribute to increasing the standard error of the $\Delta\Delta C_q$. While this example was selected as a worst-case scenario, it should illustrate that although the standard curves are seemingly well determined, it is hard to intuitively predetermine the combined effect on the standard error of $\Delta\Delta C_q$.

We note here, that no pre-averaging of the technical replicates for each concentration was done. Instead, the technical replicates where modeled as a random effect.

4.4 Simulation study

First, we present results of a simulation study for a two-sided test for the null hypothesis of a vanishing $\Delta\Delta C_q$ at a 5% significance level. We simulated 2000

datasets under both the null and alternative hypothesis with 6 samples in each case and control group and standard curves with 6 dilution steps. The effect size under the alternative was set to $\delta = 10/9$. The sample and experimental standard deviations were set to $\sigma_S = 1$ and $\sigma = 1$, respectively. The AE for the target and reference genes were set to 0.80 and 0.95, respectively.

The four discussed methods were applied to the 2×2000 datasets and the p-value testing the null hypothesis were computed. The results of these tests are summarized in Table 4 from which the FPR and TPR can be computed at the 5% cutoff. From Table 4, we see the estimated FPRs are 0.073, 0.053, and 0.083 for the efficiency corrected LMM (EC), the efficiency corrected LMM with variance adjustment using the delta method (EC&VA1), and the bootstrap, respectively. We omitted EC&VA by Monte Carlo integration here due to the computational cost and the similar results with EC&VA1 in the previous. As expected, the EC method does not control the FPR at the 5%-level. The variance adjusted estimator is consistent with controlling the FPR at the 5% level. By construction, the variance adjusted will always perform at least as good as the EC in terms of FPR. Surprisingly, the bootstrap has the worst performance in terms of FPR.

Secondly, the TPR are estimated to be 0.3825, 0.3175, 0.366 for three methods EC, EC&VA1, Bootstr., respectively. As expected, we notice that an improved FPR comes a the cost of a decreased TPR for a given statistical procedure.

The above simulations were repeated for sample sizes 4 or 8 in both case and control groups in combination with 4 or 8 dilution steps with the same simulation parameters. Fig. 4 shows the performance of the methods in terms of FPR and TPR. Each panel corresponds to a given number of samples and dilutions. In each panel the p-value cut-off is varied between 0.01, 0.05, and 0.1. Overall, we see that the EC&VA estimate is the only procedure consistent with controlling the FPR at the nominal chosen significance level. Likewise, for many dilutions, the difference between the EC and EC&VA procedures diminish as the uncertainty of the AE is relatively low. As expected a decrease in FPR corresponds to a decrease in TPR.

To gauge when the standard error of (5) is determined with adequate precision, we simulated 2×2000 datasets and computed the mean standard error of the $\Delta\Delta C_q$ for the EC and EC&VA procedures as a function of the number of dilutions and samples. We varied the number of dilutions in the range 4–9 for a number of samples in the range 4–10 with the same settings as above. Fig. 5 shows these results. As expected, increasing the number of samples or the number of dilutions yield a smaller standard error. Also unsurprising and as already seen in the applications, the differences in the standard error for the EC and EC&VA methods are very substantial for a small number of dilutions and vanish as the number of dilutions steps increase. The differences in the standard error seems to be larger under the alternative than the null hypothesis. Similar figures might also aid in designing qPCR experiments and help determine if investing in additional dilutions or samples is preferable—

Table 4: Contingency tables for the different estimators for at 5 % p-value threshold. The used estimators are the LMM with efficiency correction (EC), the LMM with EC and variance adjustment (EC&VA), and the bootstrapped LMM approach.

	EC		**EC&VA1**		**Bootstr.**	
	H_0	H_A	H_0	H_A	H_0	H_A
p-values						
$p \geq 0.05$	1854	1235	1894	1365	1834	1268
$p < 0.05$	146	765	106	635	166	732

obviously with properly chosen simulation parameters in the given context.

5 Discussion and conclusion

The commonly used efficiency corrected $\Delta\Delta C_q$ approach to analysis of qPCR data disregards the uncertainty of the estimated AE leading to increased false positive rates. As qPCR experiments are often used for validation this is highly undesirable. Our primary approach based on the statistical delta-method to approximate the variance of the efficiency adjusted $\Delta\Delta C_q$, shows that it is possible to perform statistical inference about qPCR experiments whilst more properly accounting for the AE uncertainty. We also note that the problem is not limited to the $\Delta\Delta C_q$ statistic.

The approach was used to validate that: (1) *MGST1* is differentially expressed between MM cell lines of high and low abundance of CICs and (2) analyze and study the hypothesis that miRNA-127 is differentially expressed between testicular and nodal DLBCL, and (3) illustrate the effect of a small number of dilution steps.

In the latter application, we saw a dramatic increase in the standard error of the estimate when the variance approximation was introduced, potentially leading to a change of significance for the presented dataset depending on the desired significance level. This illustrates that it is important to consider all aspects of uncertainty when conducting AE correction of qPCR experiments. Problems with uncertainty in efficiency estimates should be handled by establishing well-estimated dilution curves as argued elsewhere [5], however even in this case the presented method also allows for design guidelines for power calculations and assessing the influence of the estimated dilution curves. It is also noteworthy that model based estimation of the $\Delta\Delta C_q$ also allows for model checking by e.g. residual plots.

Lastly, we note that the algorithm [17] we used for threshold selection and C_q-value extraction in the CIC and DLBCL studies may not be optimal, cf. Ruijter et al. [26] and improvements by Spiess et al. [28], as it can be affected by the AE. Nonetheless, this has no bearing on the stated problem of this paper.

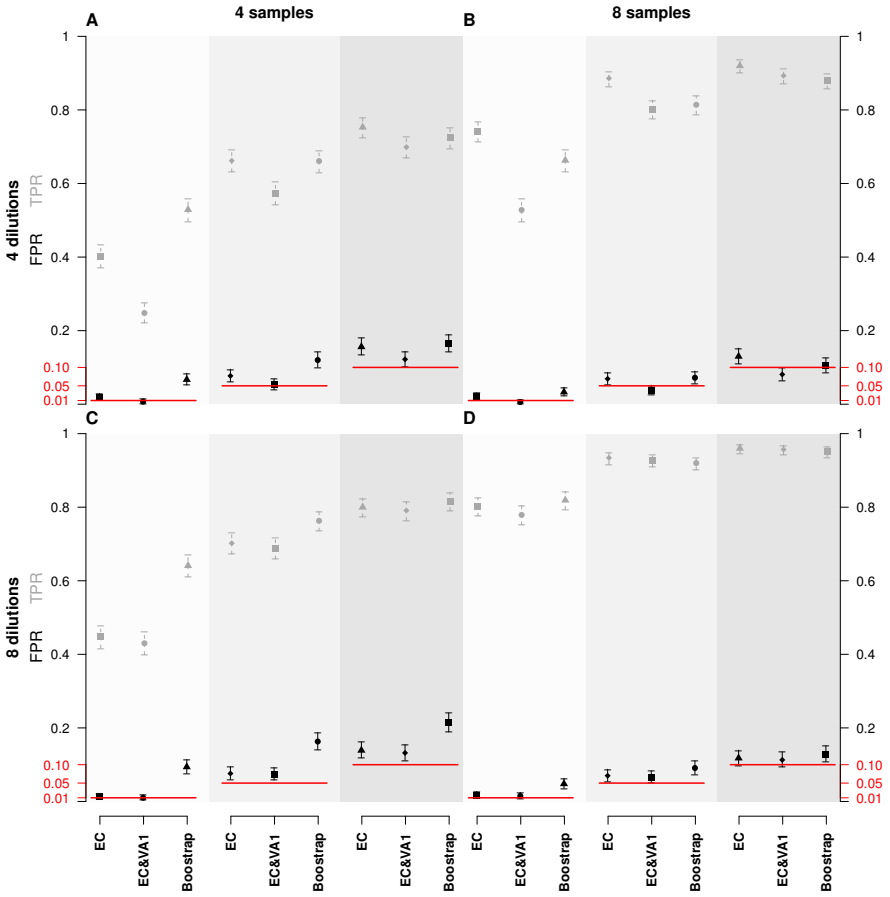


Fig. 4: Plot of the false positive rates (FPR, black) and true positive rates (TPR, grey) and their 95 % confidence intervals achieved simulation experiments for each method at various p-value cut-offs (0.05, 0.01, 0.1) shown by solid red horizontal lines. The FPR and TPR are computed completely analogous to Table 4. The rates are plotted for each combination of 4 or 8 samples with 4 or 8 fold dilution curves.

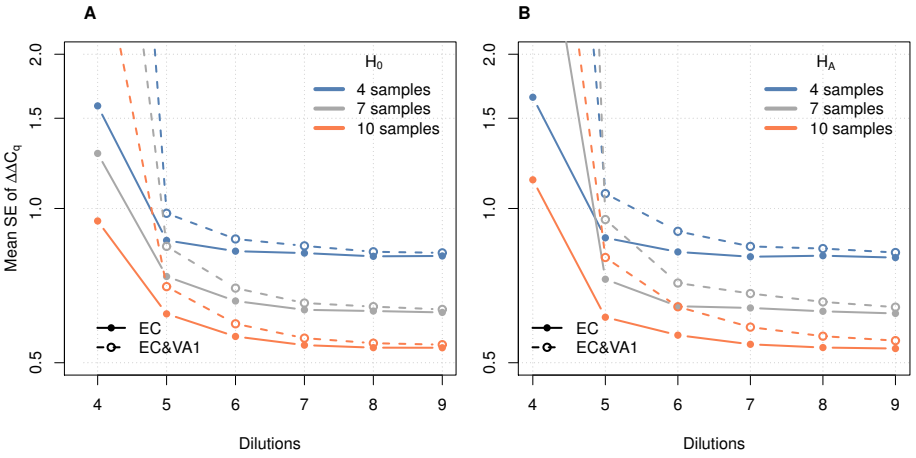


Fig. 5: The mean standard error of the $\Delta\Delta C_q$ for two methods (EC and EC&VA1) over 2000 repeated simulations under the null (panel A) and alternative hypothesis (panel B) as a function of the number of dilution steps for a different number of samples in each group.

The estimated standard error of $\Delta\Delta C_q$ is still affected in a similar manner by the uncertainty of the AE and thus too optimistic.

Despite the extensive use of qPCR, more statistical research is needed to establish qPCR more firmly as a gold standard to reliably quantify abundances of nucleic acids. Researchers analyzing qPCR experiments need to model their experiments in detail, e.g. via linear or non-linear (mixed) models, as the propagation of uncertainty needs to be carefully assessed and accounted for. This is necessary for making valid inferences and upholding the common statistical guarantees often erroneously assumed to be automatically fulfilled. We recommend the conservative and proper approach of *always* accounting for the uncertainty of the AE.

Supplementary Material and Software

The statistical analysis were done using the programming language `R v3.1.3` [24] using `lme4`. All data, R code, LaTeX, and instructions to reproduce the present paper and results are freely available at

`http://github.org/AEBilgrau/effadj/`

using `knitr`, an extension of Sweave [14, 32]. Functionality from the packages `Hmisc` [11], `lattice` (and `latticeExtra`) [27], `epiR` [30], `snowfall` [12], and `GMCM` [4], were used for tables, figures, FDR/TPR confidence intervals, parallel execution of simulations, and multivariate normal simulations, respectively.

Acknowledgments

This work was supported by the MSCNET; EU FP6; CHEPRE; Karen Elise Jensen Fonden; and the Danish Agency for Science, Technology and Innovation. The founders had no role in study design, data collection, analysis, publishing, or preparation of this paper. Poul Svante Eriksen is also thanked for his comments on the statistical methods. *Conflict of Interest*: None declared.

References

[1] L. V. Abruzzo, K. Y. Lee, A. Fuller, A. Silverman, M. J. Keating, L. J. Medeiros, and K. R. Coombes. Validation of oligonucleotide microarray data using microfluidic low-density arrays: A new statistical method to normalize real-time RT-PCR data. *Biotechniques*, 38(5):785–792, 2005. ISSN 0736-6205.

[2] C. L. Andersen, J. L. Jensen, and T. F. Ørntoft. Normalization of real-time quantitative reverse transcription-pcr data: A model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Research*, 64(15):5245–5250, 2004. ISSN 0008-5472. doi: 10.1158/0008-5472.CAN-04-0496.

[3] D. Bates, M. Maechler, B. Bolker, and S. Walker. *lme4: Linear mixed-effects models using Eigen and S4*, 2014. URL `http://CRAN.R-project.org/package=lme4`. R package v1.1-7.

[4] A. E. Bilgrau, P. S. Eriksen, J. G. Rasmussen, K. Dybkær, H. E. Johnsen, and M. Boegsted. Unsupervised clustering and meta-analysis using gaussian mixture copula models. *(Accepted for) Journal of Statistical Software*, 2015.

[5] S. A. Bustin. Why the need for qPCR publication guidelines?–the case for MIQE. *Methods*, 50(4):217–226, 2010. ISSN 1095-9130. doi: 10.1016/j.ymeth.2009.12.006.

[6] S. A. Bustin, V. Benes, J. A. Garson, J. Hellemans, J. Huggett, M. Kubista, R. Mueller, T. Nolan, M. W. Pfaffl, G. L. Shipley, et al. The MIQE guidelines: Minimum information for publication of quantitative real-time pcr experiments. *Clinical Chemistry*, 55(4):611–622, 2009. ISSN 1530-8561. doi: 10.1373/clinchem.2008.112797.

[7] K. Chen, Ying-hui, Huang, and J. long Chen. Understanding and targeting cancer stem cells: Therapeutic implications and challenges. *Acta Pharmacologica Sinica*, 34(6):732–740, 2013.

[8] B. Efron. *The Jackknife, the Bootstrap and other Resampling Plans*, volume 38. SIAM, 1982.

[9] W. J. Fu, J. Hu, T. Spencer, R. Carroll, and G. Wu. Statistical models in assessing fold change of gene expression in real-time RT -PCR experiments. *Computational Biology and Chemistry*, 30(1):21–26, 2006. ISSN 1476-9271. doi: 10.1016/j.compbiolchem.2005.10.005.

[10] R. Garzon, G. Marcucci, and C. M. Croce. Targeting micrornas in cancer: rationale, strategies and challenges. *Nature Reviews Drug Discovery*, 90: 775–779, 2010. doi: 10.1038/nrd3179.

[11] F. E. Harrell, Jr et al. *Hmisc: Harrell Miscellaneous*, 2015. URL `http://CRAN.R-project.org/package=Hmisc`. R package v3.16-0.

[12] J. Knaus. *snowfall: Easier cluster computing (based on snow)*, 2013. URL `http://CRAN.R-project.org/package=snowfall`. R package v1.84-6.

[13] I. Lefkovits and H. Waldmann. *Limiting Dilution Analysis of Cells of the Immune System*. Oxford University Press, 1999.

[14] F. Leisch. `Sweave`: Dynamic generation of statistical reports using literate data analysis. In W. Härdle and B. Rönz, editors, *Compstat 2002 — Proceedings in Computational Statistics*, pages 575–580. Physica Verlag, Heidelberg, 2002. ISBN 3-7908-1517-9.

[15] K. J. Livak and T. D. Schmittgen. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_q}$ method. *Methods*, 25(4):402–408, 2001. ISSN 1046-2023. doi: 10.1006/meth.2001.1262.

[16] M. V. Matz, R. M. Wright, and J. G. Scott. No control genes required: Bayesian analysis of qRT-PCR data. *PloS one*, 8(8):e71448, 2013.

[17] Mx3000P. *Mx3000P and Mx3005P QPCR Systems. Setup and User's Guide*. Agilent, 2013.

[18] O. Nordgård, J. T. Kvaløy, R. K. Farmen, and R. Heikkilä. Error propagation in relative real-time reverse transcription polymerase chain reaction quantification models: The balance between accuracy and precision. *Analytical biochemistry*, 356(2):182–193, 2006.

[19] M. W. Pfaffl. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Research*, 29(9):e45, 2001. ISSN 1362-4962.

[20] J. C. Pinheiro and D. M. Bates. *Mixed-effects models in S and S-PLUS*. Springer Verlag, 2000.

[21] P. Rainbow. *Making PCR*. The University of Chicago Press, 1996.

[22] X. Rao, X. Huang, Z. Zhou, and X. Lin. An improvement of the 2ˆ(-delta delta ct) method for quantitative real-time polymerase chain reaction data analysis. *Biostatistics, bioinformatics and biomathematics*, 3(3):71, 2013.

[23] X. Rao, D. Lai, and X. Huang. A new method for quantitative real-time polymerase chain reaction data analysis. *Journal of Computational Biology*, 20(9):703–711, 2013.

[24] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL http://www.R-project.org/. ISBN 3-900051-07-0.

[25] J.-L. Robertus, G. Harms, T. Blokzijl, M. Booman, D. de Jong, G. van Imhoff, S. Rosati, E. Schuuring, P. Kluin, and A. van den Berg. Specific expression of miR-17-5p and miR-127 in testicular and central nervous system diffuse large B-cell lymphoma. *Modern Pathology*, 22(4):547–555, 2009. ISSN 1530-0285. doi: 10.1038/modpathol.2009.10.

[26] J. M. Ruijter, M. W. Pfaffl, S. Zhao, A. N. Spiess, G. Boggy, J. Blom, R. G. Rutledge, D. Sisti, A. Lievens, K. De Preter, et al. Evaluation of qPCR curve analysis methods for reliable biomarker discovery: Bias, resolution, precision, and implications. *Methods*, 59(1):32–46, 2013.

[27] D. Sarkar. *Lattice: Multivariate Data Visualization with R*. Springer, New York, 2008. URL http://lmdvr.r-forge.r-project.org. ISBN 978-0-387-75968-5.

[28] A.-N. Spiess, C. Deutschmann, M. Burdukiewicz, R. Himmelreich, K. Klat, P. Schierack, and S. Rödiger. Impact of smoothing on parameter estimation in quantitative dna amplification experiments. *Clinical Chemistry*, 61 (2):379–388, 2015.

[29] J. P. Steibel, R. Poletto, P. M. Coussens, and G. J. Rosa. A powerful and flexible linear mixed model framework for the analysis of relative quantification RT-PCR data. *Genomics*, 94(2):146–152, 2009.

[30] M. Stevenson et al. *epiR: Tools for the Analysis of Epidemiological Data*, 2015. URL http://CRAN.R-project.org/package=epiR. R package v0.9-62.

[31] D. Svec, A. Tichopad, V. Novosadova, M. W. Pfaffl, and M. Kubista. How good is a pcr efficiency estimate: Recommendations for precise and robust qpcr efficiency assessments. *Biomolecular Detection and Quantification*, 3: 9–16, 2015.

[32] Y. Xie. *Dynamic Documents with* R *and* **knitr**. CRC Press, 2013. ISBN 9781482203530.

[33] Y. Yang, J. Yuan, J. Ross, J. Noel, E. Pichersky, and F. Chen. An arabidopsis thaliana methyltransferase capable of methylating farnesoic acid. *Archives of biochemistry and biophysics*, 448(1):123–132, 2006.

[34] J. S. Yuan, D. Wang, and C. N. Stewart. Statistical methods for efficiency adjusted real-time pcr quantification. *Biotechnology Journal*, 3(1):112–123, 2008. ISSN 1860-7314. doi: 10.1002/biot.200700169.

Part C

Software

Package I

GMCM: Fast Estimation of Gaussian Mixture Copula Models

Anders Ellern Bilgrau, Poul Svante Eriksen, and Martin Bøgsted

Release: `http://cran.r-project.org/package=GMCM`
Development: `https://github.com/AEBilgrau/GMCM`

Preface: The **GMCM** package is the accompanying open-source software for R described in Paper I. It greatly extends the **idr** package by Li et al. (2011) to an arbitrary number of dimensions, provides multiple algorithms for fitting the model, and implements the generalized model by Tewari et al. (2011) for unsupervised clustering. Considerable effort has gone into improving the estimation speed partly by implementing core functions in C++ via **Rcpp** and **RcppArmadillo**.

The package is extensively documented using **roxygen2** and thoroughly tested through unit tests using **testthat**. Functions not exported to the user interface also have help pages although these do not appear in the manual. They are however available in R via `help(dgmcm.loglik)` or `?dgmcm.loglik`, for example.

The manual and package vignette is available via the former link. Installation instructions and further usage examples are found by following the latter link.

Package II

DLBCLdata: Automated and reproducible download and preprocessing of DLBCL data

Anders Ellern Bilgrau and Steffen Falgreen

Development & Release: `https://github.com/AEBilgrau/DLBCLdata`

Preface: The **DLBCLdata** package for R performs automated and reproducible download and RMA preprocessing of diffuse large B-cell lymphoma gene expression datasets, freely available on the Gene Expression Omnibus (GEO) website from the National Center for Biotechnology Information (NCBI). It is partly a convenient wrapper for the **GEOquery** and **affy** packages. However, a good amount of effort went into automatically downloading and installing the custom Brainarray annotations if such non-standard preprocessing is wanted.

DLBCLdata features a number of studies that have been manually curated and checked. An overview of these studies are seen via `data(DLBCL_overview)` and used in Papers III and IV. The cleaning of the 'meta-' and clinical data is carried out by study-specific cleaning functions. The package has been written such that new studies can easily be included in the list of featured studies. An significant amount of 'manual' labour also went into ABC/GCB classifying each dataset using the Windows application by Care et al. (2013) which unfortunately did not provide a programmatic interface.

The package grew out of R code written from Paper III which uses custom brainarray ensembl-identifies. It proved itself very useful for Paper IV when we decided to use Entrez gene identifiers (as KEGG does) instead of ensembl gene ids for all studies.

While the package is oriented toward DLBCL, it should work with nearly all GEO accession numbers containing gene expression profiles on Affymetrix arrays.

Instructions for installing and using the package can be found by following the link above.

Package III

rags2ridges: Ridge estimation of precision matrices from high-dimensional data

Carel F.W. Peeters, Anders Ellern Bilgrau, and Wessel N. van Wieringen

Release:	`http://cran.r-project.org/package=rags2ridges`
Development:	`https://github.com/CFWP/rags2ridges`

Preface: The R-package **rags2ridges** performs regularized ridge estimation of (inverse) covariance matrices as discussed in van Wieringen and Peeters (2015). **rags2ridges** was expanded with a *fused*-module which encompasses the implementation of the fused ridge estimator(s) of Paper IV and many supporting functions.

To feasibly do large problems and cross-validation in the *fused* setting, much of the code base has been optimized and reimplemented in C++ via **RcppArmadillo**. This included new additional non-fused estimators which are not only faster but also more robust for extreme penalty values. This provided a speed-up in the order of a factor 100. Currently, no vignettes or usage examples exist beside the ones found in the examples of the documented functions. Extensive testing of the package functionality and interface via unit tests using **testthat** was also introduced.

ISSN (online): 2246-1248
ISBN (online): 978-87-7112-366-1

AALBORG UNIVERSITY PRESS